Christian Holberg

Exploring Irregular Dynamics: Beyond Stationarity and Continuity

PHD THESIS

this thesis has been submitted to the phd school of
the faculty of science, University of Copenhagen

Department of Mathematical Sciences
University of Copenhagen

December 2024

Christian Holberg
c.holberg@math.ku.dk
Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
2100 Copenhagen
Denmark

Thesis title:	Exploring Irregular Dynamics: Beyond Stationarity and Continuity
Supervisor:	Professor Susanne Ditlevsen University of Copenhagen
Assessment Committee:	Professor Anders Rahbek (chair) University of Copenhagen
	Professor Anastasios Magdalinos University of Southampton
	Associate Professor Blanka N. Horvath Oxford University
Date of Submission:	December 31, 2024
Date of Defense:	March 31, 2025
ISBN:	978-87-7125-239-2

This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen on the 31st of December, 2024. The PhD project was supported through grants to Susanne Ditlevsen by the Novo Nordisk Foundation (research grant NF20OC0062958) and the Independent Research Fund Denmark — Natural Sciences (research grant 9040-00215B).

Preface

The following was written during my 3 years spent as a doctoral student at the Department of Mathematical Sciences, University of Copenhagen. It contains a collection of four papers: Three published or to be published shortly and the last still a work in progress. Each of them stand as self-contained contributions, and they can be read more or less independently.

Writing a PhD has been a uniquely unpredictable experience. One that I most likely would not have been able to complete without the help of a great many people. Both academically and personally, I have been fortunate to have received a lot of support.

First and foremost, I would like to thank my supervisor, Susanne Ditlevsen. I must imagine that striking the balance between supervision and giving sufficient space is a difficult task. Nonetheless, Susanne has always managed to do exactly that, being there when guidance was needed.

Likewise, I am grateful to have worked with Cristopher Salvi, who has acted as somewhat of a second supervisor to me during this last year of my PhD. He made my stay at Imperial College London a both very pleasant and productive one.

On a personal note, I would like to thank my family for their continued support: My mom and dad, my brother Alex, his wife Asta, and my nephew Alfred. Spending time with them has been an often welcome break from academic life.

I hesitate to thank anyone individually beyond this point, as they would otherwise surely deserve. This is not due to a lack of gratitude, but only because I fear that my unreliable memory would cause me to leave someone out. Enjoying the freedom that these first few pages allow me, I am reminded of a quote of a favorite artist of mine:

> *I take everything I've already figured out with me wherever I go*
> *Mind like a moth-eaten blanket*
> *Wind whistling through the holes*
>
> — Mount Eerie, *The Gleam Pt. 3*

So, to everyone who has been but a small part of these last 3 years — friends, family, colleagues — know that I am truly thankful. Your help and support has been appreciated. Chances are, if you are reading this, this is meant for you.

Christian Holberg
December, 2024

Abstract

Contrary to the standard setting of *independent and identically distributed* (i.i.d.) data, stochastic processes may exhibit complex serial dependence structures and non-stationarity. Both of these properties complicate the statistical analysis of such processes. In this thesis, we study processes that are either non-stationary or solutions to differential equations with *event discontinuities*.

Cointegration assumes that the observed p-dimensional process is a linear mixing of k latent stationary components and $p - k$ random walks. Inference is often predicated on knowing the exact number of stationary components, i.e., the cointegration rank. Crucially, this number is unknown in practice. In the first part of this thesis, we study the asymptotic distribution of different estimators under rank uncertainty. In particular, we establish central limit theorems for reduced rank estimators in the cointegrated vector autoregressive model under misspecified rank and present a new class of weighted reduced rank estimators that are arguably more robust to rank uncertainty. We then turn to the problem of uniform inference in cointegrated vector autoregressive processes. That is, we develop asymptotic approximations for two crucial covariance statistics that are valid uniformly across a parameter space including arbitrary cointegration ranks.

In the second part of the thesis, we establish a nonlinear generalization of cointegration. We derive identification results under varying assumptions on the class of admissible mixing transformations and the non-stationary component. Then, we develop a method for estimating that stationary component based on a single discretely sub-sampled trajectory of the observable process, x_t, and show consistency under certain conditions.

Finally, in the last part of the paper, we consider the problem of deriving path-wise gradients of solutions to rough differential equations with endogenously defined discontinuities. Such discontinuities are termed *event discontinuities*. A canonical example is the spiking neuron model where an event is triggered every time the membrane potential of the neuron crosses a certain threshold upon which the potential is reset and the spike propagated to neighboring neurons. Thus, our results enable us to train spiking neuron models, where the inter-spike dynamics are governed by an SDE, using gradient-based optimization methods.

Sammenfatning

I modsætning til det klassiske tilfælde med *uafhængig og identisk fordelt* (i.i.d.) data kan stokastiske processer udvise komplekse afhængighedsstrukturer og ikke-stationaritet. Begge disse egenskaber komplicerer den statistiske analyse. I denne afhandling studerer vi processer, der enten er ikke-stationære eller løsninger til differentialligninger med *event discontinuities*.

Kointegration antager at den observerede p-dimensionelle process er en lineær blanding af k latente stationære komponenter og $p-k$ random walks. Inferens er ofte betinget af at kende det præcise antal stationære komponenter, dvs. kointegrationsrangen. Dette tal er imidlertid ukendt i praksis. I den første del af denne afhandling studerer vi den asymptotiske fordeling af forskellige estimatorer under rangusikkerhed. Vi etablerer centrale grænseværdisætninger for en klasse af estimatorer bekendt som *reduced rank estimators* i den kointegrerede vektor-autoregressive model under misspecificeret rang og præsenterer en ny klasse af estimatorer kaldet *weighted reduced rank estimators*, der er mere robuste over for rangusikkerhed. Vi vender derefter blikket mod uniform inferens i kointegrerede vektor-autoregressive processer. Vi etablerer asymptotiske approksimationer for to afgørende kovariansstatistikker, der er gyldige uniformt over et parameterområde, som inkluderer vilkårlige kointegrationsrange.

I den anden del af afhandlingen introducerer vi en ikke-lineær generalisering af kointegration. Vi udleder resultater vedrørende identifaktion under varierende antagelser over klassen af tilladte transformationer og den ikke-stationære komponent. Derefter udvikler vi en metode til at estimere den stationære komponent baseret på et enkelt diskret udsnit af den observerbare proces, x_t, og beviser konsistens under visse betingelser.

Til sidst i afhandlingen betragter vi udfordringerne ved at udlede gradienter af løsninger til såkaldte *rough differential equations* med begivenhedsdrevne diskontinuiteter. Sådanne diskontinuiteter kaldes *event discontinuities*. Et kanonisk eksempel er diskontinuerte neuronmodeller, hvor en begivenhed udløses hver gang neuronens membranpotentiale krydser en bestemt tærskel, hvorefter potentialet nulstilles, og neuronens signal propageres til naboneuronerne. Vores resultater tillader derfor kalibrering af diskontinuerte neuronmodeller, hvor dynamikken styres af en stokastisk differentialligning, ved brug af gradientbaserede optimeringsmetoder.

Contributions and Structure

In the first chapter — an introduction to the thesis — we briefly motivate our work and give a selective review of existing ideas. We then proceed with the main body of the thesis which is split into three chapters. Chapter 2 considers the problem of rank uncertainty in cointegration and contains two papers: [**WRR**] and [**STEM**]. Chapter 3 introduces a nonlinear generalization of cointegration followed by a few applications. It consists of the working paper [**STEM**]. Finally, Chapter 4 contains the paper [**SSNN**] which is concerned with deriving path-wise gradients of differential equations with endogenously defined jump discontinuities.

Chapter 2 (Beyond stationarity: Cointegration rank uncertainty)

[**WRR**] [Holberg and Ditlevsen, 2024a]. C. Holberg and S. Ditlevsen. Weighted reduced rank estimators under cointegration rank uncertainty. *Scandinavian Journal of Statistics*, 2024a. To appear.

[**UIC**] [Holberg and Ditlevsen, 2024b]. C. Holberg and S. Ditlevsen. Uniform inference for cointegrated vector autoregressive processes. *Journal of Econometrics*, 2024b. To appear.

Chapter 3 (Beyond stationarity: Nonlinear cointegration)

[**STEM**] [Holberg, 2024] C. Holberg. Stationary embeddings: A nonlinear generalization of cointegration, 2024. Working paper.

Chapter 4 (Beyond continuity: Differential equations with events)

[**SSNN**] [Holberg and Salvi, 2024] C. Holberg and C. Salvi. Exact gradients for stochastic spiking neural networks driven by rough signals. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Contents

Preface iii

Abstract iv

Contributions and Structure vii

1 Introduction 1
- 1.1 Stochastic Processes . . . 2
- 1.2 Beyond Stationarity . . . 6
- 1.3 Beyond continuity . . . 11

2 Beyond stationarity: Cointegration rank uncertainty 13

Weighted Reduced Rank Estimators Under Cointegration Rank Uncertainty 15
- 2.1 Introduction . . . 15
- 2.2 Preliminaries . . . 18
- 2.3 Asymptotic Distributions of Reduced Rank Estimators . . . 20
- 2.4 Estimation Under Rank Uncertainty . . . 25
- 2.5 Simulation study . . . 29
- 2.6 Prediction of EEG Signals . . . 32
- 2.7 Conclusion . . . 36
- 2.A Proofs . . . 38
- 2.B Multiple Lags . . . 44
- 2.C Auxiliary results . . . 45
- 2.D Simulation study . . . 47

Uniform Inference for Cointegrated Vector Autoregressive Processes 53
- 2.8 Introduction . . . 53
- 2.9 Preliminaries . . . 57
- 2.10 Asymptotic Properties . . . 59
- 2.11 Higher order VAR processes . . . 63
- 2.12 Uniform Inference . . . 65
- 2.13 Simulations . . . 72
- 2.14 Conclusion . . . 76
- 2.E Proofs . . . 77
- 2.F Confidence Regions . . . 94
- 2.G Martingale Limit Theorems . . . 95
- 2.H Gaussian Approximation . . . 98

2.I Simulations . . . 101
2.J Lag augmentation . . . 104
2.K IVX . . . 104

3 Beyond stationarity: Nonlinear cointegration 111

Stationary Embeddings: A Nonlinear Generalization of Cointegration 113
3.1 Introduction . . . 113
3.2 Stationary embeddings . . . 117
3.3 Discerning stationarity using signatures . . . 122
3.4 Estimating stationary embeddings . . . 124
3.5 Applications . . . 130
3.6 Discussion . . . 135
3.A Technical Details . . . 137
3.B Consistency . . . 140
3.C Signatures . . . 148
3.D Numerics . . . 153
3.E Experiments . . . 159

4 Beyond continuity: Differential equations with events 165

Exact Gradients for Stochastic Spiking Neural Networks Driven by Rough Signals 167
4.1 Introduction . . . 167
4.2 Related work . . . 169
4.3 Stochastic spiking neural networks as Event SDEs . . . 171
4.4 Training stochastic spiking neural networks . . . 176
4.5 Conclusion . . . 179
4.A Càdlàg rough paths . . . 181
4.B Proof of Theorem 2 . . . 187
4.C Kernel methods . . . 194
4.D Forward sensitivities for SLIF network . . . 196
4.E Experiments . . . 198

Bibliography 201

1 Introduction

We will be concerned with the difficulties that arise when working with data where observations tend to depend on each other. This is often the case in time series data where observations are obtained from the same individual or system at different points in time. In order to make inference feasible, some minimum set of assumptions must be imposed on the data generating mechanism. Often, we can get away with treating the data as if it were i.i.d. under an appropriate such set of assumptions. This means that we can simply proceed "as per usual". However, caution is warranted. The data at hand need not conform to our assumptions and many subtle violations will generally lead to faulty conclusions. To illustrate this point, let us consider the simple phenomenon of *spurious regression* [Granger and Newbold, 1974].

Example 1.0.1. Suppose we are given two univariate time series $(x_t)_{t\geq 1}$ and $(y_t)_{t\geq 1}$. One may posit the linear regression model $y_t = \beta x_t + \varepsilon_t$, where ε_t is some sequence of error terms, and then employ the well-known machinery of *ordinary least squares regression*. That is, we obtain an estimate of β via the formula

$$\hat{\beta} = \frac{\sum_{t=1}^{n} x_t y_t}{\sum_{t=1}^{n} x_t^2}.$$

If both x_t and y_t are stationary autoregressive processes and the errors are, for example, i.i.d., all the classical asymptotic results hold. In particular, $\hat{\beta} - \beta$ converges at rate $\sqrt{n}$ in distribution to a normal distribution centered at 0 enabling us to construct confidence intervals by using the standard procedure. In other words, we do not really have to worry about the dependency structure of our data. The situation quickly changes, however, if both x_t and y_t are non-stationary. Indeed, let x_t and y_t be independent random walks initialized at $x_0 = y_0 = 0$ implying that $\beta = 0$. Then, $\hat{\beta}$ no longer converges in probability to 0, but instead converges in distribution to some non-degenerate random variable [Phillips, 1986], i.e., in general we would find that $\hat{\beta} \neq 0$ even though the two time series are independent. ♠

The concept of spurious regression is by now well-studied in the context of *persistent* time series (see, e.g., [Lee et al., 2005, Phillips, 2009, Tu and Wang, 2022]), but the preceding example still highlights what can go wrong when handling time series data.

In this thesis, we will consider systems which are *irregular* in the sense that they give rise to non-standard dependency structures that make straightforward application of classical statistical or machine learning methods difficult. Broadly, the work can be split into two strands. The first strand deals with non-stationary processes similar to the example above. As it turns out, it is in fact possible to obtain valid results from

least squares regression even when x_t and y_t exhibit random-walk-like behavior as long as they are *cointegrated*. In fact, this is more or less the definition of cointegration. The second strand is of a different nature. Here, we study systems that can be described by certain kinds of differential equations with jump discontinuities. Such jumps can occur because some exogenous noise affects the system in a discontinuous manner, but they could also occur due to some endogenously triggered mechanism. Think, for example, of a bouncing ball whose velocity abruptly changes every time it hits the ground. The jump is then triggered by its own position. We call such endogenous jumps *events*. Our primary motivation for studying differential equations with events is their applicability to neuron models where spikes are triggered by the membrane potential.

To start, we shall introduce a few concepts that are essential for both strands. Throughout, continuous time stochastic processes and, specifically, those arising from *stochastic differential equations* (SDEs), will play an important role either appearing as distributional limits or as the primary object of interest.

1.1 Stochastic Processes

Throughout this thesis, a stochastic process refers to a collection of random variables taking values in a measurable space and indexed by some set $\mathcal{T}$. For all our purposes, we can think of $\mathcal{T}$ as representing time. That is, for some $T \geq 0$ denoting the possibly infinite final time, we assume that either $\mathcal{T} = (0 \leq t_1 < \cdots < t_n < T)$ or $\mathcal{T} = [0, T)$. The former case corresponds to a discrete stochastic process or a *time series* while the latter is a continuous time stochastic process. For the sake of being precise, let $(E, \mathcal{E})$ be a measurable state space and consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. A stochastic process is then simply a collection of E-valued random variables $(x_t)_{t\in\mathcal{T}}$ each defined on the same underlying probability space, $(\Omega, \mathcal{F}, \mathbb{P})$. In most of the following, it suffices to take $E = \mathbb{R}^p$ meaning that we shall mainly be concerned with Euclidean stochastic processes. When we are sure to avoid confusion, we simply write x_t or even x to refer to the whole process. One may view x as a map from Ω to $E^{\mathcal{T}}$, the set of all functions from $\mathcal{T}$ to E. The law of x_t is then simply given by $\mathbb{P} \circ x^{-1}$, the push-forward of $\mathbb{P}$ under x.

Usually, the easiest way to construct a stochastic process is to specify its *finite-dimensional distributions*. That means, for any $n \geq 1$ and $t_1, \ldots, t_n \in \mathcal{T}$, we specify a probability measure $\mu_{t_1,\ldots,t_n}$ on E^n. As long as these finite-dimensional distributions are compatible (which essentially means that they are well-behaved under marginalization), the Kolmogorov extension theorem guarantees the existence of a unique stochastic process whose law, when restricted to $\{t_1, \ldots, t_n\}$, agrees with $\mu_{t_1,\ldots,t_n}$ for any $n \geq 1$ and $t_1, \ldots, t_n \in \mathcal{T}$. At this point we note that stochastic processes subsume the i.i.d. setting. Indeed, a sequence of i.i.d. random variables is simply a stochastic process whose finite-dimensional distributions all factorize to powers of the same identical probability measure. Allowing for dependency between subsequent observations of the sequence, the finite dimensional distributions become more involved.

For a thorough introduction to stochastic processes, we refer to, e.g., Chapter 2 of

Rogers and Williams [2000].

1.1.1 From discrete to continuous time

In practice, we will never encounter data that is continuous in time for the simple reason that we can only measure a system a countable number of times. As a result, it would appear that continuous time processes serve little purpose for any pragmatically minded data scientist. This conclusion, however, would be a mistake. Continuous time processes serve a range of useful purposes both on the level of modelling, but also as convenient asymptotic approximations of many objects arising from time series data. On the level of modelling, SDEs are an important tool. They are an extension of ordinary differential equations where the state is also affected by some external highly irregular and noisy control (usually Brownian motion). We shall describe these in more detail in the subsequent section. For now, let us give two examples of how continuous time processes naturally appear in many asymptotic results concerning time series data.

Example 1.1.1. Let $x_t = x_{t-1} + \varepsilon_t$ be a univariate random walk with $\varepsilon_t \sim N(0,1)$ i.i.d. and define, for every $n \geq 1$, the sample covariance, $\hat{\sigma}_n^2$, given by

$$\hat{\sigma}_n^2 := \frac{1}{n} \sum_{t=1}^{n} x_t^2.$$

The variance of x_t increases with time. In fact, $\mathrm{Var}(x_t) = t$. Naturally, then, the sample covariance diverges as well. However, when normalized by n, it converges in distribution to the Riemann integral of the square of a standard Brownian motion. To be precise, as $n \to \infty$,

$$\frac{\sigma_n^2}{n} \to_d \int_0^1 w_t^2 dt,$$

where w_t is a standard Brownian motion. ♠

Example 1.1.2. In addition to the sample covariance of Example 1.1.1, define now, for every $n \geq 1$, the sample cross-covariance, $\hat{\gamma}_n$, given by

$$\hat{\gamma}_n := \frac{1}{n} \sum_{t=1}^{n} x_t \varepsilon_t.$$

One might expect $\hat{\gamma}_n$ to converge to 0, but this is not the case. Instead, it converges in distribution to the Itô integral of w_t against itself. That is, as $n \to \infty$,

$$\hat{\gamma}_n \to_d \int_0^1 w_t dw_t.$$

♠

Both of these asymptotic results, or at least their multivariate counterparts, play an important role in *cointegration*, a concept that we shall define shortly. We study

generalizations of these results particularly in [**UIC**]. As stated here, Example 1.1.1 and 1.1.2 are little more than an application of Donsker's invariance principle (see, e.g., Theorem 8.2 in Rogers and Williams [2000]) which itself has applications even in the realm of i.i.d. data by yielding a central limit theorem for empirical distribution functions.

1.1.2 Stochastic differential equations

Apart from the typical *drift term*, a stochastic differential equation has an additional *diffusion term* corresponding to some exogenous noisy control. For two vector fields, $\mu : \mathbb{R}^p \to \mathbb{R}^p$ and $\sigma : \mathbb{R}^p \to L(\mathbb{R}^q, \mathbb{R}^p)$, we say that $x_t \in \mathbb{R}^p$ is the solution to the SDE with drift μ, diffusion σ, and initial condition $a \in \mathbb{R}^p$ if

$$dx_t = \mu(x_t)dt + \sigma(x_t)dw_t, \quad x_0 = a \tag{1}$$

where $w_t \in \mathbb{R}^q$ is the exogenous noise. Often times, w_t is taken to be the standard q-dimensional Brownian motion. The differential notation is best understood in its integral form, that is, integrating both side of the equality with the second integral on the right-hand side defined as a stochastic integral. Multiple choices are possible with the two most common by far being the *Itô* or *Stratonovich* integrals which give rise, respectively, to Itô or Stratonovich SDEs. Although, we point out that when w_t is a general semimartingale (that might be discontinuous), there also exists another interpretation of (1) as a *Marcus* SDE [Marcus, 1978, 1981]. We return to this latter interpretation in [**SSNN**].

Existence and uniqueness of the different kinds of SDEs can be established under conditions similar to the ones familiar from the study of ODEs. For an introduction to SDEs, we refer to Oksendal [2013] and for a more thorough exposition, see Karatzas and Shreve [1991].

Interpreted as an Itô SDE, if x_t is the solution to (1), at each point in time and for $\delta > 0$ small enough, $x_{t+\delta}$ is approximately given by the drift at the current state, $\mu(x_t)\delta$, plus the diffusion at the current state applied to the incremental change of the control, $\sigma(x_t)(w_t - w_{t+\delta})$. This is also known as the Euler approximation. Importantly, the noisy term directly affects the state and therefore also the future dynamics. Note that this is conceptually different from an ODE with noisy measurements where noise is only introduced a posteriori.

Example 1.1.3. Perhaps the simplest non-trivial SDE is the Ornstein-Uhlenbeck process which is the solution to (1) with $\mu(x) = \Pi x$ and $\sigma(x) = \sigma$ for $\Pi \in \mathbb{R}^{p\times p}$ and $\sigma \in \mathbb{R}^{p\times q}$. This is one of the few SDEs that can be solved analytically, i.e., we have an expression for x_t that can be written completely in terms of the driving noise, w_t. In particular,

$$x_t = ae^{t\Pi} + \int_0^t e^{(t-s)\Pi}\sigma dw_s.$$

When w_t is the standard Brownian motion, it follows that x_t is a Gaussian process. We also note that, for any fixed increment $\delta > 0$, x_t satisfies the recursive relation $x_{t+\delta} = e^{\delta\Pi}x_t + \varepsilon_t$ where ε_t is a centered Gaussian vector with covariance matrix given by

$\int_0^\delta e^{t\Pi}\Sigma e^{t\Pi}dt$ and $\Sigma = \sigma\sigma^T$. In other words, when discretizing an Ornstein-Uhlenbeck process, one obtains a vector autoregressive process of order 1. In general, if we replace the linear drift by any arbitrary nonlinear function, the discretization of the resulting SDE will no longer necessarily be an autoregressive process. However, the Euler approximation will. ♠

We emphasize that it is quite rare to be able to solve (1) analytically. More often than not, we have to resort to numerical methods to approximate the solution. This also means that, contrary to the Ornstein-Uhlenbeck process, we usually do not have a nice way to express the distribution of a discrete sub-sample. Two questions then naturally spring to mind if we want to learn the parameters of the SDE based on data. Which objective should we minimize to find our estimates? And secondly, how do we compute the gradient of this objective with respect to the model parameters? An answer to the first question is to use the likelihood when possible and otherwise to use an approximation (see, e.g., Pilipovic et al. [2024a,b]). Another way is to take a sort of non-parametric approach and compare the observed discrete sample paths with the ones generated by the SDE using some statistical discrepancy or score function. This is the approach taken, for example, in Issa et al. [2023b] to train *neural SDEs* where the authors use a score function based on the *signature maximum mean discrepancy*. This discrepancy uses the signature which is a concept from rough path theory and can be thought of as a feature map for path-valued data. We discuss the signature as well as other useful concepts from rough path theory in the following section.

There are multiple answers to the second question. Let us focus on the problem of obtaining path-wise gradients of solutions of (1) with respect to the initial condition, a, since this enables us to compute the gradients of any objective used in practice. Note that other parameters, call them $\theta \in \mathbb{R}^{p_\theta}$, can be included in the initial condition by augmenting (1) with the additional constant state $d\theta_t = 0$, $\theta_0 = \theta$. The term *path-wise* refers to the fact that we are interested in a random variable, $g \in \mathbb{R}^{p\times p}$, such that, for almost every $\omega \in \Omega$, it holds that $g(\omega) = \partial_a x_T(\omega)$. Taking for granted the existence of path-wise gradients, one approach to finding them does so by differentiating the numerical solution of the SDE. That is, since most numerical SDE solvers are composed of differential primitives, we can simply backpropagate through the solver to obtain the gradients of the sample paths. This is here referred to as the *discretize-then-optimize* approach. A second approach computes the gradients by solving a second related SDE usually known as the *adjoint equation*. One can show that $g = \lambda_0$ where λ_t is the solution to the adjoint equation. Of course, the adjoint equation is usually as hard to solve as the original SDE, so one would use a numerical solver also for this. This is the *optimize-then-discretize* approach. The gradients obtained in each of these two ways align if we use a *reversible solver*. We have borrowed the terminology from Kidger [2021] where much more information on this topic can be found.

1.1.3 Some perspectives from rough path theory

Rather recently, a new way to interpret the SDE (1) has emerged seeming more natural if one is coming at it from the direction of controlled differential equations. Indeed, one

of the key contributions of rough path theory is that it allows us to define differential equations driven by very irregular paths (think Brownian sample paths) with randomness acting as a fully separate component. To be more precise, on a set of measure 1, we can define path-wise solutions to (1) as solutions to a *rough differential equation* (RDE), that is, given a sample of a Brownian rough path, we can solve the differential equation deterministically. This, of course, should be contrasted against the classical interpretation from stochastic calculus where the Itô integral — and, hence, the solution to the SDE — is only defined probabilistically. To construct a Brownian rough path, we need to enhance the sample paths with an additional level two tensor (or matrix) for which multiple choices are possible. Two common choices correspond to Itô and Stratonovich SDEs respectively. In fact, for SDEs driven by general semimartingales, including discontinuous ones, it is possible to define a rough path such that solutions to the corresponding RDE coincide a.s. with the SDE (regardless of the interpretation). A more thorough introduction to RDEs driven by càdlàg rough paths is given in [**SSNN**]. An excellent exposition of rough path theory can be found in Friz and Victoir [2010] or, for a more gentle introduction to the topic, in Lyons et al. [2007].

Another important object from rough path theory is the *signature*. In essence, for a path of bounded variation, x_t, the signature is an infinite sequence of tensors where the terms are given by iterated integrals of the form

$$\int \cdots \int_{0<u_1<\cdots<u_k<t} dx_{u_1} \otimes \cdots \otimes dx_{u_k} \in (\mathbb{R}^p)^{\otimes k}.$$

Among other things, signatures serve as universal feature maps [Lemercier et al., 2021] and characterize the law of stochastic processes [Chevyrev and Oberhauser, 2022]. They are intimately related with Taylor expansions of controlled differential equations. Assuming that the path x_t takes values in some Hilbert space, the signature will also be Hilbert space valued. Taking inner products of signatures results in the signature kernel for which kernel tricks exist [Salvi et al., 2021b, Király and Oberhauser, 2019] circumventing the need to compute the signature explicitly. The signature kernel maximum mean discrepancy then serves as a natural metric for probability distributions on path space.

1.2 Beyond Stationarity

As touched upon above, if the data is *stationary*, we may get lucky in that inference proceeds as if the data was i.d.d. Usually this requires some additional assumptions such as ergodicity or certain mixing conditions, but these are not as important for now and will follow for many classes of processes if just stationarity can be established. In words, a process is stationary if it behaves the same no matter when we measure it. There are multiple equivalent definitions of stationarity. We shall opt for the following: For $\mathcal{T} = [0, \infty)$ and any $\tau > 0$, we define the shift operator $\theta_\tau : E^{\mathcal{T}} \to E^{\mathcal{T}}$ taking $f(\cdot)$ into $f(\tau + \cdot)$. A stochastic process $x \in E^{\mathcal{T}}$ is then stationary exactly when the law of x and $\theta_\tau(x)$ agree for all $\tau > 0$. This definition is easily carried over to time series;

simply replace $[0, \infty)$ with the natural numbers. Sometimes it makes sense to work with a weaker notion of stationarity. We say that a process is *weakly stationary* if the mean and variance does not depend on time.

Let us now return to the discretization of the Ornstein-Uhlenbeck process from Example 1.1.3 or, more generally, to the vector autoregressive process of order 1 (VAR(1)), $x_t = \Gamma x_{t-1} + \varepsilon_t$. We have already seen that x_t is non-stationary when $\Gamma = I$, i.e., when x_t is a random walk. Note that this is equivalent to saying that all eigenvalues of Γ are exactly 1. On the other hand, it is well-known that, if Γ has eigenvalues strictly bounded in length by 1 (or if Π has eigenvalues with negative real part in the continuous time analog), then x_t is stationary and ergodic. Now a natural question is what happens if only some eigenvalues are equal to 1. In this case, the process will still be non-stationary and exhibit persistent behavior similar to that of a random walk, but there is a sense in which it is also partly stationary; It will be *cointegrated*.

Being a little imprecise for now, we say that a time series, x_t, is *cointegrated of order* $0 \leq k \leq p$ if it is non-stationary, but $\Delta x_t := x_t - x_{t-1}$ is stationary and there exists k linearly independent linear combinations of the coordinates of x_t that are stationary. For the sake of exposition, we here restrict ourselves to cointegration for time series, but one may just as well define it for continuous time processes. See, for example, Kessler and Rahbek [2001, 2004]. In the cointegration literature stationarity is usually synonymous with weak stationarity. The cointegration rank of a p-dimensional VAR(1) process like the one above is exactly equal to p minus the number of eigenvalues equal to 1 (also called unit roots). Cointegration was first introduced in the seminal work of Engle and Granger [1987] with much of the theory for VAR processes then developed in Johansen [1988, 1995]. Cointegrated VAR processes are best represented using the vector error correction model (VECM),

$$\Delta x_t = \Pi x_{t-1} + \varepsilon_t \tag{2}$$

where $\Pi = \Gamma - I$ and $\Delta x_t = x_t - x_{t-1}$. Γ having $p - k$ unit roots is then exactly equivalent to the rank of Π being k and estimating Π can be done using reduced rank regression [Anderson, 2002b, Johansen, 1995]. Since Π is of rank $k \leq p$, we can find two matrices $\alpha, \beta \in \mathbb{R}^{p \times k}$ such that $\Pi = \alpha\beta^T$. As it turns out, under some minor technical assumptions, the cointegration relations are given by the columns of β, i.e., $y_t := \beta^T x_t \in \mathbb{R}^k$ is stationary.

Alternatively, taking a less parametric angle, cointegration can also be formulated using the following model

$$x_t = (A_1, A_2)(y_t, z_t) \tag{3}$$

where $A = (A_1, A_2) \in \mathbb{R}^{p \times p}$ is invertible with $A_1 \in \mathbb{R}^{p \times k}$, $A_2 \in \mathbb{R}^{p \times (p-k)}$. A constitutes the mixing transformation and $y_t \in \mathbb{R}^k$ is the stationary latent component and $z_t \in \mathbb{R}^{p-k}$ is non-stationary with stationary first differences. The cointegration relations are then given by the first k columns of the inverse mixing A^{-1}. Under this formulation, the problem is also sometimes referred to as *stationary subspace analysis* (SSA) [Von Bünau et al., 2009]. We note that for this model, as well as for the previous one, there is an issue of identifiability. For example, in the VECM, for any $Q \in \mathbb{R}^{k \times k}$ invertible, we can define $\tilde{\alpha} = \alpha Q$ and $\tilde{\beta} = \beta Q^{-1}$ so that $\tilde{\alpha}\tilde{\beta}^T = \Pi = \alpha\beta^T$. This issue is often resolved

by imposing a specific normalization. For example, one can choose the following upper triangular parameterization of A which is the framework employed in, e.g., Phillips [1991],

$$A = \begin{pmatrix} I_k & B_2 \\ 0 & I_{p-k} \end{pmatrix}.$$

Furthermore, in either model, there is the question of picking the appropriate cointegration rank. Having access only to the observed process x_t, this is a non-trivial problem. In practice, the most common approach estimates the rank by performing a nested sequence of reduced rank tests stopping at the first rank for which the null hypothesis can no longer be rejected. Johansen [1988] developed likelihood ratio tests based on Gaussian error terms for the null hypothesis $H_0 : \text{rank}(\Pi) = k$ against either of the alternatives $H_A : \text{rank}(\Pi) = k + 1$ or $H_A : \text{rank}(\Pi) = p$ in the VECM model.

Thus, cointegration presents a middle ground allowing us to handle non-stationary data. Much of this thesis is devoted to studying cointegrated processes. Our work on cointegration can roughly be split into two parts. The first part looks at how uncertainty around the cointegration rank can affect inference. The second part gives a suggestion for how to define a nonlinear extension of cointegration.

1.2.1 Rank uncertainty

Since we are estimating the cointegration rank from data, it should be clear that a certain probability of error is to be expected. For example, if x_t is high-dimensional, there is a large probability that we stop our sequential testing procedure too early [Stærk-Østergaard et al., 2023, Onatski and Wang, 2018]. On the other hand, Γ may have roots that are arbitrarily close to unity and therefore impossible to detect with finite samples. This can be modelled using the so-called local-to-unity setup [Phillips, 2009].

As such, it becomes essential to establish the asymptotic behavior of estimators under rank uncertainty. In [**WRR**] we establish central limit theorems for reduced rank estimators of Π under two regimes: 1) the chosen rank is greater than the true rank and 2) the chosen rank is less than the true rank. We compare the two asymptotic distributions to the classical case where one assumes that the rank is correctly specified. Similar results have been shown in the i.i.d. setting in Anderson [2002a]. As one would have perhaps expected, a bias is incurred if the rank is underestimated while the asymptotic variance increases in the opposite case. We explore how these results manifest themselves in estimation and prediction. We also show that the reduced rank estimators are part of a wider family of estimators which we call the *weighted reduced rank estimators*. Based on empirical results, we argue that the flexibility of this new class of estimators is beneficial in settings with rank uncertainty. The weighted reduced rank estimator can be seen as a weighted average of the individual reduced rank estimators of each rank from $k = 0$ to $k = p$. Surprisingly little work has been done on model averaging for cointegrated VAR processes. The only reference we have been able to find doing work in this direction is Lieb and Smeekes [2017].

As a corollary to the results in [**WRR**], we find that the least squares estimator of Γ,

i.e.,

$$\hat{\Gamma} = \left(\sum_{t=1}^{n} x_t x_t^\mathsf{T}\right)^{-1} \sum_{t=1}^{n} x_t x_{t-1}^\mathsf{T},$$

is consistent regardless of the true rank. This is, of course, not a new result. But even then, the coast is not clear since the asymptotic distribution still depends on the number of unit roots. In fact, even in the univariate case where Γ is a scalar and therefore only has one eigenvalue, it was shown in Elliott [1998] that the statistical analysis is heavily dependent on whether Γ is close to 1. The need arises for methods of inference that are robust to slight deviations from the unit root assumption. The sensible way to formalize this is in terms of uniform inference, that is, we seek methods with asymptotic level holding uniformly over a range of suitable parameters. This is the content of [**UIC**].

The local-to-unity regime models Γ as a sequence of parameters converging to I from below at rate n. Specifically, it assumes that $\Gamma = I - C/n$ for some diagonal $C \in \mathbb{R}^{p\times p}$ with positive real entries. Under this assumption the asymptotic distribution of the least squares estimator, $\hat{\Gamma}$, is different from either the stationary or the random walk ($\Gamma = I$) case. In the local-to-unity regime, the asymptotic distribution will depend on the latent parameter C which is not consistently estimable. Mikusheva [2007] show that, for $p = 1$, varying the scalar C between 0 and ∞, we obtain a family of distributions that interpolate the classical normal asymptotics of the stationary univariate autoregressive process and the non-standard asymptotics of the random walk. This results in another family of distributions that, asymptotically, can uniformly approximate the least squares estimator for Γ arbitrarily close to 1. Mikusheva [2007] then used this result to formally verify which of the supposedly robust methods of inference did in fact yield uniformly valid confidence intervals (or statistical tests).

Of course, the story is much more involved in the multivariate setting if not only for the fact that we can no longer work with the simple parameter space $\Gamma \in [1-\delta, 1]$.[1] We note that, since

$$\hat{\Gamma} - \Gamma = \left(\sum_{t=1}^{n} x_t x_t^\mathsf{T}\right)^{-1} \sum_{t=1}^{n} x_t \varepsilon_t^\mathsf{T},$$

the asymptotic distribution of (an appropriately normalized version of) $\hat{\Gamma} - \Gamma$ is determined by the asymptotic behavior of the sample covariance matrices

$$S_{xx} = \frac{1}{n}\sum_{t=1}^{n} x_t x_t^\mathsf{T}, \quad S_{x\varepsilon} = \frac{1}{n}\sum_{t=1}^{n} x_t \varepsilon_t^\mathsf{T}.$$

The reader may recognize these quantities as multivariate generalizations of the sample covariance and cross-covariance from Example 1.1.1 and 1.1.2. In [**UIC**], we show that, as was the case for $p = 1$, we can use the local-to-unity regime to obtain families of distributions that, asymptotically, uniformly approximate S_{xx} and $S_{x\varepsilon}$ over a parameter

[1] Here $\delta > 0$ is some small constant bounding Γ away from -1. For $p \geq 1$, we only allow unit roots (or eigenvalues of Γ with modulus 1) that are exactly 1 since this would otherwise imply that x_t is *seasonally cointegrated*.

space that allows for the unit roots of Γ to be arbitrarily close to 1. Thus, we do for the VAR process what Mikusheva [2007] did for the AR process.

1.2.2 Nonlinear cointegration

Coming up with a sensible notion of non-linear cointegration has proven difficult. We refer to Tjøstheim [2020] for a review of some recent advances where the authors argue that the problem, or at least a part of it, is that classical linear cointegration deals only with non-stationarity in the form of integrated processes. Crucially, this property is not preserved under arbitrary nonlinear transformation and therefore is ill-suited for a nonlinear formulation of cointegration. We argue that one benefits from taking a wider view and considering all forms of non-stationarity. That is, a nonlinear generalization of cointegration requires that the stationary component be *strictly stationary* and that the non-stationary components are just that, arbitrarily non-stationary.

As with the linear case, there are several ways to approach the problem. The formulation one chooses ultimately depends on the problem at hand. Perhaps the most common way to look at it is in the form of nonlinear regression, $x_t = f(z_t) + y_t$, where x_t and z_t are non-stationary and y_t is stationary [Park and Phillips, 1999, 2001]. Another common way is to consider nonlinear extension of the VECM [Bec and Rahbek, 2004]. None of these suggestions are particularly satisfying, though, not in the least because they are unnecessarily restrictive. For a truly general definition, we may take as our starting point the model (3) which is reminiscent of linear *blind source separation* (BSS).

To obtain a nonlinear definition of cointegration is then as simple as replacing the invertible linear map A with a general invertible smooth map $d : \mathbb{R}^p \to \mathbb{R}^p$. Another subtle point is that, as mentioned earlier, we initially only require that y_t is stationary — in the proper sense of that word — and that z_t is non-stationary. In other words, no other special assumptions are made about the distribution of y_t and x_t. Now, learning the latent components from an observation of x_t is hard. First, as in the linear case, there is the question of identifiability. In the linear case we know that the stationary component is identifiable up to invertible linear transformations while the non-stationary component is not identifiable. This latter statement is essentially due to the fact that adding a stationary process to a non-stationary one will yield another non-stationary process. In [**STEM**] we show that a similar result holds in the nonlinear case. Specifically, we can identify the stationary component up to invertible smooth transformations. If $e = (e_1, e_2) = d^{-1}$ so that $y_t = e_1(x_t)$, this means that $\tilde{y}_t = f(x_t) \in \mathbb{R}^k$ is stationary if, and only if, f is equal to $h \circ e_1$ for some invertible smooth map h.

Our goal is to estimate e based on the observed process x_t keeping in mind that our identifiability results would then imply that we can only do so up to the resulting equivalence class. We take inspiration from Schell and Oberhauser [2023] and construct an objective, call it φ, that discriminates stationary from non-stationary stochastic processes. Estimating e is then as easy as minimizing this statistic over a suitably flexible function class. Of course, there are many small practical details to address when working with actual data. We introduce a function based on signature kernels similar to Issa and Horvath [2023] that acts as a test statistic for the null hypothesis of stationarity and

discuss practical numerical implementations of the resulting procedure.

1.3 Beyond continuity

Returning now to the general SDE in (1), in many applications it is sensible to allow the driving noise, w_t, to be discontinuous. Think, for example, of a neuron in a network where the effect of unobserved spike trains may be modelled as Poisson processes. Another example could be a financial time series that is subject to sudden big shocks in the market. Now, there are many ways to interpret the SDE (1) (or the corresponding RDE) in the presence of jumps, but what all of them have in common is that the solution, x_t, will be jump-discontinuous as well. In fact, its discontinuity points align exactly with those of w_t. We call such jumps *exogenous* since they are known a priori, i.e., their timing is completely determined by the driving noise w_t. We contrast these types of jumps with another kind of *endogenous* jump that is triggered by the state of the solution itself. In order to determine the timing of these jumps, we would need to first solve (1). One way to formalize this is to augment the differential equation with an additional object, $\mathcal{E} : \mathbb{R}^p \to \mathbb{R}$, known as the *event function*. The endogenous jumps, which we call *event times*, are then simply solutions to the stopping time problem $\inf_{t>s}\{\mathcal{E}(x_t) = 0\}$. To describe what happens once an event is triggered, we can define a *transition function*, $\mathcal{T} : \mathbb{R}^p \to \mathbb{R}^p$. Upon triggering an event, the transition function then maps the current state of the solution to a new state after which x_t again behaves according to the dynamics given by (1). We call systems like these *event SDEs* (or event RDEs) and they are rigorously defined in [**SSNN**]. Phrased in a slightly different way, event SDEs fall under the umbrella of stochastic hybrid systems [Lygeros and Prandini, 2010] for which a fair amount of literature exists, but we note that the extension to RDEs is new.

Consider now the differential equation (1) as a RDE, that is, w_t is (possibly a sample of) some rough path. For $T \geq 0$, we can define the flow map $\Phi : \mathbb{R}^p \mapsto \mathbb{R}^p$ that maps an initial condition a to the value of the solution of (1) at time T. It is well-known that this is a smooth function even when w_t is discontinuous (or càdlàg, to be precise) [Chevyrev and Friz, 2019]. The crucial question, then, is whether a similar result can be shown to hold for event RDEs. When w_t is smooth it has been known for some time that the answer to this question is positive. See, e.g., Corner et al. [2019, 2020] and, later, Chen et al. [2018] where the results are restated in the context of neural differential equations. Specifically, one can use the implicit function theorem to prove that the event times are differentiable from which smoothness of the flow follows as well. This strategy, unfortunately, does not work if w_t is rough. This is essentially due to the fact that w_t is then nowhere differentiable. What we show in [**SSNN**] is that, using a limiting argument that is common in rough path theory, we can extend the result from the smooth case to the rough case.

This result has important implications, for example, for calibration of spiking neural networks. Using an extension of the signature kernel maximum mean discrepancy to càdlàg paths, we show how that this enables training networks of stochastic spiking neurons by minimizing the discrepancy with gradient descent.

2 Beyond stationarity: Cointegration rank uncertainty

This chapter contains the following two papers:

[**WRR**] [Holberg and Ditlevsen, 2024a]. C. Holberg and S. Ditlevsen. Weighted reduced rank estimators under cointegration rank uncertainty. *Scandinavian Journal of Statistics*, 2024a. To appear.

[**UIC**] [Holberg and Ditlevsen, 2024b]. C. Holberg and S. Ditlevsen. Uniform inference for cointegrated vector autoregressive processes. *Journal of Econometrics*, 2024b. To appear.

Throughout we are concerned with vector autoregressive processes of order $p \geq 1$ (VAR(p) processes). In particular, we shall consider a time series $(x_t)_{t\in\mathbb{N}}$ generated recursively by

$$x_t = \Gamma x_{t-1} + \varepsilon_t$$

where $\Gamma \in \mathbb{R}^{p\times p}$ and ε_t is some sequence of errors. In [**WRR**] we shall assume that the errors are i.i.d., but it is entirely possible to allow for serial dependence in the errors as in [**UIC**] where ε_t is assumed to be a martingale difference sequence. The nice thing about VAR processes is that cointegration is directly characterized by the eigenstructure of Γ. In particular, if, for every eigenvalue, $\lambda \in \mathbb{C}$, of Γ, it holds that $|\lambda| \leq 1$ with equality if and only if $\lambda = 1$, then the cointegration rank of x_t is exactly p minus the number of eigenvalues equal to 1. We call such eigenvalues unit roots. In [**WRR**] we consider what happens if inference is conducted under a misspecified cointegration rank. On the other hand, in [**UIC**] we focus solely on the least squares estimator, but let the parameters vary freely. In particular, we develop uniformly (across the parameter space) valid asymptotic approximations of the estimator.

Weighted Reduced Rank Estimators Under Cointegration Rank Uncertainty

Christian Holberg, Susanne Ditlevsen

Abstract

Cointegration analysis was developed for non-stationary linear processes that exhibit stationary relationships between coordinates. Estimation of the cointegration relationships in a multi-dimensional cointegrated process typically proceeds in two steps. First the rank is estimated, then the auto-regression matrix is estimated, conditionally on the estimated rank (reduced rank regression). The asymptotics of the estimator is usually derived under the assumption of knowing the true rank. In this paper, we quantify the asymptotic bias and find the asymptotic distributions of the cointegration estimator in case of misspecified rank. Furthermore, we suggest a new class of weighted reduced rank estimators that allow for more flexibility in settings where rank selection is hard. We show empirically that a proper choice of weights can lead to increased predictive performance when there is rank uncertainty. Finally, we illustrate the estimators on empirical EEG data from a psychological experiment on visual processing.

2.1 Introduction

2.1.1 Motivation

Consider a p-dimensional autoregressive process Y_t of order 1 (AR(1)) defined by a vector error correction model (VECM)

$$\Delta Y_t = \Pi Y_{t-1} + Z_t \tag{2.1.1}$$

where $\Delta Y = Y_t - Y_{t-1} \in \mathbb{R}^p$, Π is the $p \times p$ autoregression matrix of fixed coefficients of rank $r \leq p$, and $Z_1, Z_2 \ldots$ are i.i.d. p-dimensional random vectors of mean zero.

In standard low-dimensional problems, the typical procedure to determine r is based on sequential likelihood-ratio tests [Johansen, 1995]. The test statistics do not follow any standard distributions, the critical values depend on p and they need to be calculated numerically. Currently, critical values are available for dimension $p \leq 11$. This can be overcome by bootstrap methods. However, it is nontrivial to keep control over the type I

error and the sequential testing can lead to severe bias, especially when the dimension p of model (2.1.1) increases [Stærk-Østergaard et al., 2023, Onatski and Wang, 2018]. Once the rank is fixed, a reduced rank regression [Anderson, 2002b] is performed assuming that this rank is in fact the true rank.

In settings where rank estimation is hard or where the rank is fixed a priori, questions arise regarding properties of the estimator. One such question is how to characterize the asymptotic behaviour of reduced rank estimators where the rank is fixed at a value not necessarily equal to the true rank. Furthermore, treating the rank obtained from the sequential testing approach as fixed in the subsequent analysis neglects the added uncertainty. Thus, there is a need for more flexible estimators that take the uncertainty into account. We suggest a weighted average of reduced rank estimators where ranks are weighted depending on the supporting evidence in the data. All the classical reduced rank estimators are special cases of this more general class of estimators.

2.1.2 Literature Review

A lot of work has been done on the asymptotic behaviour of reduced rank estimators under the true rank or assuming wrongly full rank $r = p$ (see, for example, Johansen [1988, 1995], Anderson [2002b] for the time series setting and Izenman [1975], Anderson [1999] for the i.i.d setting). Less work has been done under the assumption of a misspecified rank, that is, cases where the rank of the reduced rank estimator is not equal to the true rank of Π. For the i.i.d. setting we refer to [Anderson, 2002a].

The work closest to ours are Bernstein and Nielsen [2019] and Cavaliere et al. [2012]. Bernstein and Nielsen [2019] study the asymptotic distribution of likelihood ratio tests on the cointegration matrix in the rank deficient case. Their results differ from ours in that they consider the cointegration matrix instead of the full autoregression matrix Π. Also, while their focus is on hypothesis testing specifically, our main results concern the estimator. Lemma 1 in Cavaliere et al. [2012] contains a result on the consistency (in the sense of convergence to certain "pseudo parameters") of the reduced rank estimators when the rank is underestimated. The authors do not, however, provide an asymptotic distribution. See also Remark 1 below.

Model averaging and weighted estimators for cointegrated VAR processes has also received little attention. See Koop et al. [2006] for a review of Bayesian approaches. Hansen [2010] deals with model averaging for one-dimensional processes with potential unit root. Lieb and Smeekes [2017] is most closely related to our approach, but they only consider one family of weights and their main concern is inference.

One approach that is often used when facing rank uncertainty (especially in high-dimensional problems) is penalization. For a review of reduced rank estimators with a fixed rank, using different types of penalizations, see Levakova and Ditlevsen [2023].

2.1.3 Our contribution

There are three main contributions of the present paper. The first is to determine the asymptotic distribution of the reduced rank estimator of Π under misspecified rank r.

This has important statistical implications since there is no guarantee of determining even closely the true rank from finite sample sizes, especially for large p. To this end, we first consider a linearly transformed version of the system given in (2.1.1) in which the "stationary" and "random walk" directions are separated. We show for this transformed system that the reduced rank estimator is consistent but has increased variance when the rank is overestimated, compared to a correctly specified rank. In the original coordinates, that is, when inverting the transformation, this translates to an increase of variance in the "random walk" direction. However, asymptotically this increase vanishes on the $\sqrt{T}$ scale (see 4). If the rank is underestimated, an asymptotic bias is introduced. The bias depends on the sizes of the eigenvalues of a certain eigenvalue problem. The main results are Theorem 2 and 3. Especially the proof of Theorem 3 is interesting. It relies on the delta method applied to a central limit theorem for a specific covariance matrix (Lemma 3) requiring us to carry out some novel computations involving matrix derivatives.

Our second contribution is the introduction of a new class of estimators which we call *weighted reduced rank estimators*. We show how the classical reduced rank estimators are special cases of this class and how Theorem 2 and 3 determine the asymptotic behaviour for any particular weight $w \in [0,1]^p$. We argue why taking rank uncertainty into account is appropriate and show empirically in a simulation experiment how the predictive capabilities of the weighted reduced rank estimators outperform the classical reduced rank estimator based on pre-selected rank.

Our third contribution is the application of the new estimator on an experimental data set of Electroencephalography (EEG) measurements in a visual response study. We then compare the performance with the fixed rank estimators. In this study, $p = 59$ with relatively small sample sizes, which is a typical setting where the rank is not well determined. We show that the smaller the sample size, the better the weighted reduced rank estimators perform compared to the fixed rank estimators, for any fixed rank, measured on mean square prediction error. This is important in many neurobiological studies, where the data dimension is high but sample size is restricted to a small time interval if a response to a stimulus is of interest.

We emphasize that our analysis pertains to settings where the full autoregression matrix Π is of interest. In particular, we do not focus on estimation of the cointegration matrix $\beta \in \mathbb{R}^{d\times r}$ where $\Pi = \alpha\beta^T$. Examples where estimating Π is relevant are for forecasting or impulse response analysis. In a VAR(1) model, the impulse responses at different periods are given by powers of Π [Lütkepohl, 2005]. Our results therefore have direct relevance for impulse response estimation. For example, the asymptotics of different estimators of Π^k under varying ranks can be derived directly from our main results and the delta method.

2.1.4 Organization

The paper is organized as follows. In Section 2.2 the model and assumptions for cointegration are presented. In Section 2.3 the asymptotic distribution under correctly specified rank is recalled and the main results are presented, namely the asymptotic distributions under misspecified rank. In Section 2.4 we introduce the weighted reduced rank

estimators. Section 2.5 consists of two simulation experiments to verify our asymptotic results and compare the different estimators. Section 2.6 compares different weighted reduced rank estimators on a dataset of EEG signals and Section 2.7 concludes. All the proofs are presented in Section 2.A. In the Appendix some auxiliary results are given. We show how the framework extends to processes of higher lags and give some further details regarding the simulations.

2.1.5 Notation

I_k denotes the k-dimensional identity matrix. Transposition is denoted by ^{T}. Convergence in distribution is denoted by $\to_w$ and convergence in probability by $\to_p$. The Frobenius norm is denoted by $||\cdot||_F$. For a matrix $A \in \mathbb{R}^{n\times m}$ with $n \geq m$ of full column rank m, we write $A_\perp$ to denote the $n \times (n-m)$ matrix of full column rank $(n-m)$ such that $\text{span}(A)^\perp = \text{span}(A_\perp)$. If A is positive definite we write $A^{\frac{1}{2}}$ for the unique positive definite matrix satisfying $A^{\frac{1}{2}}A^{\frac{1}{2}} = A$. The vectorization operator is written as vec.

2.2 Preliminaries

Let $\{Y_t\}_{t=1}^\infty$ be defined by (2.1.1). Autoregressive processes of higher order, VAR(d) with $d > 1$, are briefly treated in Appendix 2.B. Assume that $Z_1, Z_2, ...$ are i.i.d. of mean zero, with covariance matrix $\Sigma_Z := \mathbb{E}(Z_t Z_t^T)$, and a bounded fourth moment. Furthermore, assume that the process satisfies the usual cointegration assumption for some $0 \leq r \leq p$:

Assumption 1. *The polynomial $z \mapsto |(1-z)I_p - \Pi z|$ has $n = p - r$ unit roots and all other roots are outside the unit circle.*

This assumption implies that the rank of Π is $p - n = r$. Thus, we can decompose Π into two matrices $\alpha, \beta \in \mathbb{R}^{p\times r}$ of rank r such that $\Pi = \alpha\beta^T$. Let $\alpha_\perp$ and $\beta_\perp$ be orthogonal complements of α and β. This leads to the second condition that is usually assumed when working with cointegrated AR-processes [Johansen, 1995].

Assumption 2. *The $n \times n$ matrix $\alpha_\perp^T\beta_\perp$ is non-singular.*

Under these assumptions Granger's representation theorem [Engle and Granger, 1987, Johansen, 1991] states that Y_t is integrated of order 1 ($I(1)$) and cointegrated of rank r. The cointegration relations are given by $\beta^T Y_t$. That is, Y_t exhibits random walk like behaviour with ΔY_t and $\beta^T Y_t$ being stationary. Now define $Q = (\beta, \alpha_\perp)^T$ and note that

$$Q^{-1} = \left(\alpha(\beta^T\alpha)^{-1}, \beta_\perp(\alpha_\perp^T\beta_\perp)^{-1}\right).$$

Then, with $X_t = QY_t$, $U_t = QZ_t$, and

$$\Gamma = Q\Pi Q^{-1} = \begin{pmatrix} \beta^T\alpha & 0 \\ 0 & 0 \end{pmatrix},$$

we get the Q-transformed version of model (2.1.1),

$$\Delta X_t = \Gamma X_{t-1} + U_t. \tag{2.2.2}$$

We have effectively split up the original process Y_t into a stationary part and a random walk part. In particular, if X_{1t} denotes the first r components of X_t and X_{2t} the last n components, we have the following relations

$$\Delta X_{1t} = \beta^T \alpha X_{1t-1} + U_{1t} \tag{2.2.3}$$
$$\Delta X_{2t} = U_{2t}. \tag{2.2.4}$$

We shall first study estimators of Γ from observations $X_0, X_1, \ldots, X_T$ and then transfer the results to the original parameter of interest, Π. The reason for taking this small detour is that it will give more clarity to the limiting behaviour of different parts of the estimator corresponding to either the random walk or the stationary part of the process.

Before describing the asymptotics of the estimators we need some results regarding the cross-covariances. Specifically, define the empirical cross-covariances

$$S_{XX} = \frac{1}{T}\sum_{t=1}^{T} X_{t-1}X_{t-1}^T, \quad S_{UX} = \frac{1}{T}\sum_{t=1}^{T} U_t X_{t-1}^T,$$

$$S_{\Delta XX} = \frac{1}{T}\sum_{t=1}^{T} \Delta X X_{t-1}^T, \quad S_{\Delta X \Delta X} = \frac{1}{T}\sum_{t=1}^{T} \Delta X \Delta X^T,$$

and the covariance matrix $\Sigma_U = \mathbb{E}(U_t U_t^T) = Q\Sigma_Z Q^T$. We use the following block matrix notation: For a $p \times p$ matrix M, let M_{11} denote the top left $r \times r$ block, M_{22} the bottom right $n \times n$ block, and M_{12} and M_{21} the two off-diagonal blocks. For notational convenience, we sometimes use a superscript instead. We implicitly assume that all the limits considered in the following sections are for $T \to \infty$. For the stationary processes, X_{1t-1} and ΔX_t, the law of large numbers yields

$$S_{XX}^{11} \to_p \Sigma_X^{11} = \sum_{s=0}^{\infty} (I_r + \beta^T\alpha)^s \Sigma_U^{11} (I_r + \alpha^T\beta)^s$$
$$S_{\Delta X \Delta X} \to_p \Sigma_{\Delta X} = \begin{pmatrix} \Sigma_U^{11} + \beta^T\alpha\Sigma_X^{11}\alpha^T\beta & \Sigma_U^{12} \\ \Sigma_U^{21} & \Sigma_U^{22} \end{pmatrix}.$$

To study the asymptotics of the random walk part of the process we introduce a standard p-dimensional Brownian motion initiated at 0 denoted by $\{W_s\}_{s\in[0,1]}$. We are now ready to present the crucial Lemma. A proof can be found in e.g. Lemma 7.1 in Lütkepohl [2005].

Lemma 1. *Define the $n \times p$ matrix $D = (0, I_n)$ and the random $r \times p$ matrix $V^T := (V_{11}^T, V_{21}^T)$ satisfying* $\mathrm{vec}V \sim \mathcal{N}(0, \Sigma_X^{11} \otimes \Sigma_U)$. *The following converge jointly:*

$$T^{-1}S_{XX}^{22} \to_w D\Sigma_U^{\frac{1}{2}} \left(\int_0^1 W_s W_s^T \, ds \right) \Sigma_U^{\frac{1}{2}} D^T =: B \tag{2.2.5}$$

$$\begin{pmatrix} S_{UX}^{12} \\ S_{UX}^{22} \end{pmatrix} \to_w \Sigma_U^{\frac{1}{2}} \left(\int_0^1 W_s \, dW_s^T \right)^T \Sigma_U^{\frac{1}{2}} D^T =: \begin{pmatrix} J_{12} \\ J_{22} \end{pmatrix} \tag{2.2.6}$$

$$T^{\frac{1}{2}} \begin{pmatrix} S_{UX}^{11} \\ S_{UX}^{21} \end{pmatrix} \to_w V = \begin{pmatrix} V_{11} \\ V_{21} \end{pmatrix}. \tag{2.2.7}$$

Furthermore,

$$S_{XX}^{11} \to_p \Sigma_X^{11} \tag{2.2.8}$$

$$S_{\Delta X \Delta X} \to_p \Sigma_{\Delta X}. \tag{2.2.9}$$

A direct consequence of the above Lemma and summation by parts is that $S_{\Delta XX}^{12} \to_p -\Sigma_U^{12}$ (see section 3.1 in Anderson [2002b]). Thus, we have fully uncovered the asymptotic behaviour of $S_{\Delta XX}$ as well. Indeed, we can write

$$\begin{pmatrix} S_{\Delta XX}^{11} & S_{\Delta XX}^{12} \\ S_{\Delta XX}^{21} & S_{\Delta XX}^{22} \end{pmatrix} = \begin{pmatrix} \beta^T \alpha S_{XX}^{11} & \beta^T \alpha S_{XX}^{12} \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} S_{UX}^{11} & S_{UX}^{12} \\ S_{UX}^{21} & S_{UX}^{22} \end{pmatrix}.$$

We have already established that the top right block converges in probability to $-\Sigma_U^{12}$ and limits for the remaining three blocks follow easily from Lemma 1. Note that, whereas $S_{\Delta XX}^{11}$, $S_{\Delta XX}^{12}$, and $S_{\Delta XX}^{21}$ converge in probability, $S_{\Delta XX}^{22} = S_{UX}^{22}$ converges only weakly. From this it also follows that S_{XX}^{12} and S_{XX}^{21} are bounded in probability whence $T^{-\frac{1}{2}} S_{XX}^{12}, T^{-\frac{1}{2}} S_{XX}^{21} \to_p 0$.

2.3 Asymptotic Distributions of Reduced Rank Estimators

With Lemma 1 in our arsenal, we are ready to study the asymptotic behaviour of estimators of Γ. In particular, we shall focus on the standard cointegration estimators [Johansen, 1995]. This is a collection of estimators that can be obtained by solving a generalized eigenvalue problem. We consider

$$|S_{X\Delta X}(S_{\Delta X \Delta X})^{-1} S_{\Delta XX} - \hat{\lambda} S_{XX}| = 0$$

and order the solutions in decreasing order, $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ... \geq \hat{\lambda}_p$. With $\hat{\Lambda} := \text{diag}(\hat{\lambda}_1, ..., \hat{\lambda}_p)$, denote by $\hat{G}$ the $p \times p$ matrix solving

$$S_{X\Delta X}(S_{\Delta X \Delta X})^{-1} S_{\Delta XX} \hat{G} = S_{XX} \hat{G} \hat{\Lambda}, \tag{2.3.10}$$

$$\hat{G}^T S_{XX} \hat{G} = I_p. \tag{2.3.11}$$

In column vector notation we write $\hat{G} = (\hat{g}_1, ..., \hat{g}_p)$. For any $m_1 \times m_2$ matrix M we shall write $M^{:k}$ for the $m_2 \times k$ matrix consisting of the first $k \leq m_2$ columns of M. Keeping in line with our previous block matrix notation, we write $\hat{G}_{11}$ and $\hat{G}_{22}$ for the top left $r \times r$ block and the bottom right $n \times n$ block of $\hat{G}$ respectively and $\hat{G}_{21}$ and $\hat{G}_{12}$ for the off diagonal blocks. The reduced rank estimators are given by

$$\hat{\Gamma}_k = S_{\Delta XX} \hat{G}^{:k} \left(\hat{G}^{:k} \right)^T \tag{2.3.12}$$

for $k = 0, ..., p$ (where $k = 0$ is trivial). $\hat{\Gamma}_k$ is called the reduced rank estimator of Γ for rank k. One can show that $\hat{\Gamma}_k$ is the maximum likelihood estimator of Γ under Gaussian errors when the data generating process is given by (2.2.2) and the rank of Γ is fixed

at k [Johansen, 1995]. In our case the true rank is $0 \leq r \leq p$ and the reduced rank estimator for a correctly specified rank is therefore $\hat{\Gamma}_r$. Another special case is the least squares estimator $\hat{\Gamma}_{\text{LS}}$ which is, in fact, equal to $\hat{\Gamma}_p$.

For future reference, we also define the (appropriately rescaled) population versions of $\hat{\lambda}$ and $\hat{G}$. We let $\lambda_1 \geq \cdots \geq \lambda_p$ be the ordered solutions to

$$\left| \begin{pmatrix} \Sigma_X^{11} \alpha^T \beta \left(\Sigma_{\Delta X}^{-1}\right)_{11} \beta^T \alpha \Sigma_X^{11} & 0 \\ 0 & J_{22} \end{pmatrix} - \lambda \begin{pmatrix} \Sigma_X^{11} & 0 \\ 0 & B \end{pmatrix} \right| = 0 \tag{2.3.13}$$

with $G = (g_1, \ldots, g_p)$ the corresponding eigenvectors normalized so that

$$G^T \begin{pmatrix} \Sigma_X^{11} & 0 \\ 0 & B \end{pmatrix} G = I_p.$$

The block diagonal structure implies that almost surely G_{12} and G_{21} are 0 with G_{11} and G_{22} solving the two seperate eigenvalue problems defined by the diagonal blocks in (2.3.13). This furthermore implies that the first r eigenvalues and eigenvectors are deterministic while the last n eigenvalues and eigenvectors are random. We let $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$.

It makes sense to distinguish between three different situations and study them separately. First, the reduced rank estimator where the true rank is given a priori. In this case we include exactly enough information and the resulting estimator is optimal, among the estimators considered here, in the following sense: For all $0 \leq k \leq p$ for which $\hat{\Gamma}_k$ is consistent, $\hat{\Gamma}_r$ has the lowest asymptotic variance.

Knowing the number of cointegrating relations, however, is often unrealistic. This leads us to consider the estimators $\hat{\Gamma}_{k_1}$ when $0 \leq k_1 < r$ and $\hat{\Gamma}_{k_2}$ when $r < k_2 \leq p$. The former has underestimated rank and we will show that it is asymptotically biased, but under some circumstances the bias might be small enough to make it preferable in a bias-variance trade-off. The latter has overestimated rank and we will show that it is consistent, but its variance is inflated when compared to $\hat{\Gamma}_r$. We first recall the known limiting behaviour of $\hat{\Gamma}_r$ since this serves as an illustrating case and highlights many of the ideas involved in the study of the other two cases. We then derive the limiting behavior of the estimators with misspecified ranks, which is the first main contribution of this paper.

2.3.1 Correctly Specified Rank

We start with a result due to Anderson [2002b]. The statement of the Theorem as well as the proof are essentially the same as in Anderson [2002b]. A proof can be found in Section 2.A.

Theorem 1. *Define* $\tilde{J}_{12} := (J_{12} - \Sigma_U^{12}(\Sigma_U^{22})^{-1} J_{22})$ *and let* $\text{rank}(\Gamma) = r$. *Then,*

$$\begin{pmatrix} T^{\frac{1}{2}}(\hat{\Gamma}_r^{11} - \Gamma_{11}) & T(\hat{\Gamma}_r^{12} - \Gamma_{12}) \\ T^{\frac{1}{2}}(\hat{\Gamma}_r^{21} - \Gamma_{21}) & T(\hat{\Gamma}_r^{22} - \Gamma_{22}) \end{pmatrix} \to_w \begin{pmatrix} V_{11}(\Sigma_X^{11})^{-1} & \tilde{J}_{12} B^{-1} \\ V_{21}(\Sigma_X^{11})^{-1} & 0 \end{pmatrix}, \tag{2.3.14}$$

where J_{12}, J_{22} *and* B *are defined in Lemma 1.*

Note that the rate of convergence for the right two blocks is $o_P(T^{-1})$ contrary to the usual reduced rank regression setting of independent observations where the rate of convergence is $o_P(T^{-\frac{1}{2}})$ for all blocks. This is because $T\hat{G}_{21}^{:r}(\hat{G}_{11}^{:r})^T$ and $T^2\hat{G}_{21}^{:r}(\hat{G}_{21}^{:r})^T$ are convergent, where $\hat{G}$ is defined in (2.3.10)–(2.3.11), as can be seen in the proof.

2.3.2 Overestimated Rank

Let the true rank of Π be $0 \leq r < p$. We are interested in the reduced rank estimator $\hat{\Gamma}_{r+m}$ with $r < r+m \leq p$. The above results for $\hat{\Gamma}_r$ suggest that this estimator is consistent and with a limiting behaviour somewhat close to that of $\hat{\Gamma}_r$ depending on m. To tackle this problem, we first analyze the asymptotics of the last n columns of $\hat{G}$. Unfortunately, we cannot directly adopt the methods from the previous section, but in much the same way we start with (2.3.10) and (2.3.11).

Consider equation (2.3.11) in block matrix notation. Using that $(I_r, T^{\frac{1}{2}}I_n)\hat{G}$ is bounded in probability and that $T^{\frac{3}{4}}\hat{G}_{21}$ converges in probability to 0 (see proof of Theorem 1), we find that $\hat{G}_{21}^T S_{XX}^{21}\hat{G}_{12}$, $\hat{G}_{11}^T S_{XX}^{12}\hat{G}_{22}$, and $\hat{G}_{21}^T S_{XX}^{22}\hat{G}_{22}$ are $o_P(T^{-\frac{1}{4}})$. The top-right block of (2.3.11) then reduces to

$$\hat{G}_{11}^T S_{XX}^{11}\hat{G}_{12} + o_P(T^{-\frac{1}{4}}) = 0$$

so that $\hat{G}_{12} = o_P(T^{-\frac{1}{4}})$. By an analogous argument we get $\hat{G}_{22}S_{XX}^{22}\hat{G}_{22} = I_n + o_P(T^{-\frac{1}{2}})$ and therefore $T\hat{G}_{22}\hat{G}_{22}^T = (T^{-1}S_{XX}^{22})^{-1} + o_P(T^{-\frac{1}{2}})$. Note that the least squares estimator is given by

$$\hat{\Gamma}_{\mathrm{LS}} = S_{\Delta XX}(S_{XX})^{-1} = S_{\Delta XX}\hat{G}\hat{G}^T = \hat{\Gamma}_p$$

corresponding to the case where $m = n$. Thus, we have obtained an asymptotic distribution for the least squares estimator, albeit in a slightly indirect way. This will also be a consequence of the following more general result.

Theorem 2. *Assume that* $\mathrm{rank}(\Gamma) = r < p$ *and* $1 \leq m \leq n = p - r$. *Then,*

$$\begin{pmatrix} T^{\frac{1}{2}}(\hat{\Gamma}_{r+m}^{11} - \Gamma_{11}) & T(\hat{\Gamma}_{r+m}^{12} - \Gamma_{12}) \\ T^{\frac{1}{2}}(\hat{\Gamma}_{r+m}^{21} - \Gamma_{21}) & T(\hat{\Gamma}_{r+m}^{22} - \Gamma_{22}) \end{pmatrix} \to_w \begin{pmatrix} V_{11}(\Sigma_X^{11})^{-1} & \tilde{J}_{12}B^{-1} + \tilde{J}_{22}P_m \\ V_{21}(\Sigma_X^{11})^{-1} & J_{22}P_m \end{pmatrix} \tag{2.3.15}$$

where $\tilde{J}_{22} := \Sigma_U^{12}(\Sigma_U^{22})^{-1}J_{22}$ *and* $P_m = G_{22}^{:m}(G_{22}^{:m})^T$ *with* G *defined in* (2.3.13).

It is instructive to compare the limiting distribution in (2.3.15) with (2.3.14). What effectively happens when inflating the rank is that we are including columns of $\hat{G}$ that are not relevant. This leads to an increased variance as illustrated by the terms $\tilde{J}_{22}G_{22}^{:m}(G_{22}^{:m})^T$ and $J_{22}G_{22}^{:m}(G_{22}^{:m})^T$. The higher m is, the more the variance increases. For small m compared to p, there might not be any major issues. In line with our intuition, it is thus advisable to get as close as possible to the true rank. Setting $m = 0$ corresponds to dropping all columns of $G_{22}^{:m}$ and we end up with (2.3.14). For the least squares estimator the above expression simplifies somewhat. Indeed, $G_{22}^{:n} = G_{22}$ and thus $G_{22}^{:n}(G_{22}^{:n})^T = B^{-1}$. Plugging this into (2.3.15) yields $\tilde{J}_{12}B^{-1} + \tilde{J}_{22}G_{22}^{:n}(G_{22}^{:n})^T = J_{12}B^{-1}$ and $J_{22}G_{22}^{:n}(G_{22}^{:n})^T = J_{22}B^{-1}$.

2.3.3 Underestimated Rank

For finite samples, we might just as well underestimate the true rank, especially if one chooses the rank using the sequential testing approach that is usually applied in practice. We now consider $\hat{\Gamma}_m$ for $0 \leq m < r$. It is clear that the estimator will not be consistent so all we can hope for is that the asymptotic bias is small in certain situations. Before computing this bias and giving the main theorem of this section, we need an extra assumption on the generalized eigenvalues in (2.3.13).

Assumption 3. *The first r generalized eigenvalues in (2.3.13) are simple, i.e., $\lambda_1 > ... > \lambda_r$.*

This assumption is of a technical nature. It is needed for the smoothness results given in Lemma 2 and 3 in the Appendix. To the extent in which we apply Lemma 2, it is actually sufficient to assume $\lambda_m > \lambda_{m+1}$. It is clear that this assumption is necessary since otherwise we would not be able to distinguish between the asymptotic eigenvectors. Whether we need all the first r eigenvalues to be simple, however, is questionable. We hypothesize that Theorem 3 holds without this assumption, but this would require a different proof since the current proof relies on the delta method, which in turn requires sufficient smoothness of a certain map of the generalized eigenvectors.

It immediately follows from Lemma 2 in Appendix 2.C and the proof of Theorem 1 that $\hat{G}_{11}^{:m}(\hat{G}_{11}^{:m})^T \to_p G_{11}^{:m}(G_{11}^{:m})^T$. Furthermore, we know that $\hat{\Gamma}_m^{11} = \beta^T\alpha S_{XX}^{11}\hat{G}_{11}^{:m}(\hat{G}_{11}^{:m})^T + o_P(1)$. Then, since $\beta^T\alpha = \beta^T\alpha\Sigma_X^{11}G_{11}G_{11}^T$, we find that the asymptotic bias is given by

$$\hat{\Gamma}_m^{11} - \Gamma_{11} \to_p \beta^T\alpha\Sigma_X^{11}(G_{11}G_{11}^T - G_{11}^{:m}\left(G_{11}^{:m}\right)^T) =: b_m. \tag{2.3.16}$$

We see that the asymptotic bias increases as eigenvalues are excluded and the bias is larger for larger eigenvalues. In practice this means that we only incur a small bias when underestimating the rank if the eigenvalues $\lambda_{m+1}, ..., \lambda_r$ are small.

We obtain the following asymptotic distribution of the reduced rank estimator when the rank is underestimated. A proof can be found in Section 2.A.

Theorem 3. *Assume that $0 \leq m < r = \text{rank}(\Gamma)$ and $\lambda_1 > \cdots > \lambda_r$. Let κ_{ijkl} be the joint cumulant of $U_{t,i}, U_{t,j}, U_{t,k}$, and $U_{t,l}$ and assume furthermore that $\kappa_{ijkl} = 0$ for all $1 \leq i, j, k, l \leq p$. Let $\tilde{V}^T = (\tilde{V}_{11}^T, \tilde{V}_{21}^T)$ be a random matrix such that $\text{vec}(\tilde{V}) \sim \mathcal{N}(0, \xi\Xi\xi^T)$, where Ξ is defined in (2.A.5) and ξ is defined in (2.A.7). Then,*

$$\begin{pmatrix} T^{\frac{1}{2}}(\hat{\Gamma}_m^{11} - \Gamma_{11} - b) & T(\hat{\Gamma}_m^{12} - \Gamma_{12}) \\ T^{\frac{1}{2}}(\hat{\Gamma}_m^{21} - \Gamma_{21}) & T(\hat{\Gamma}_m^{22} - \Gamma_{22}) \end{pmatrix} \to_w \begin{pmatrix} \tilde{V}_{11} & C_m\tilde{J}_{12}B^{-1} \\ \tilde{V}_{21} & 0 \end{pmatrix} \tag{2.3.17}$$

where $C_m = \beta^T\alpha\Sigma_X^{11}G_{11}^{:m}(G_{11}^{:m})^T(\beta^T\alpha)^{-1}$. The covariance matrix of $\tilde{V}_{21}$ is equal to $G_{11}^{:m}(G_{11}^{:m})^T\Sigma_X^{11}G_{11}^{:m}(G_{11}^{:m})^T \otimes \Sigma_U^{22}$.

From $G_{11}^{:r}(G_{11}^{:r})^T = (\Sigma_X^{11})^{-1}$ it follows that $C_r = I_r$. Comparing (2.3.17) with (2.3.14) we see that the variances of the top right and bottom left blocks are reduced. The bottom right block also converges in probability to 0. Comparison of the top left blocks

is more involved. We could not find a straightforward answer to prefer one over the other. Interestingly enough, simulations suggest that the variance may even increase in certain parts when lowering the rank m.

When $m = r$ the expression for ξ simplifies to the one derived in Theorem 1, see Section 2.A.

2.3.4 Asymptotics in the Original Coordinates

Recall that X_t defined by (2.2.2) was a transformation of Y_t defined by (2.1.1) into coordinates where the stationary and random-walk parts of the process are separated. Our original parameter of interest was Π. We now discuss how to derive central limit theorems for a family of estimators of Π analogous to those discussed above. In particular, we define for $0 \leq k \leq p$ the matrices $\hat{L}_k = Q^T \hat{G}_k$ and the estimators $\hat{\Pi}_k = S_{\Delta YY} \hat{L}^{:k} (\hat{L}^{:k})^T$. Then $\hat{L}^{:k}$ solves

$$S_{Y\Delta Y}(S_{\Delta Y\Delta Y})^{-1} S_{\Delta YY} \hat{L}^{:k} = S_{YY} \hat{L}^{:k} \hat{\Lambda}^{:k:k}$$
$$(\hat{L}^{:k})^T S_{YY} \hat{L}^{:k} = I_k$$

where $\Lambda^{:k:k} = \operatorname{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_k)$ and the solutions to $|S_{Y\Delta Y}(S_{\Delta Y\Delta Y})^{-1} S_{\Delta YY} - \hat{\lambda} S_{YY}| = 0$ are the same as those for the Q-transformed cross-covariances. The columns of $\hat{L}$ are thus the generalized eigenvectors for the generalized eigenvalue problem given by $S_{Y\Delta Y}(S_{\Delta Y\Delta Y})^{-1} S_{\Delta YY}$ and S_{YY}. Furthermore, we have $\hat{\Pi}_k = Q^{-1} \hat{\Gamma}_k Q$ and, by definition, $\Pi = Q^{-1} \Gamma Q$. Consequently, a central limit theorem for $\hat{\Pi}_k$ is easily obtained from Theorems 1, 2, and 3.

Theorem 4. *Assume that* $0 \leq r = \operatorname{rank}(\Pi) \leq p$. *Then, if* $r \leq k \leq p$,

$$T^{\frac{1}{2}} \operatorname{vec}(\hat{\Pi}_k - \Pi) \to \mathcal{N}(0, \beta(\Sigma_X^{11})^{-1} \beta^T \otimes \Sigma_Z).$$

If we furthermore assume that $\lambda_1 > \ldots > \lambda_r > 0$ *and that* $\kappa_{ijkl} = 0$ *for all* $1 \leq i, j, k, l \leq p$, *then, for* $0 \leq k < r$,

$$T^{\frac{1}{2}} \operatorname{vec}(\hat{\Pi}_k - \Pi - \tilde{b}) \to \mathcal{N}(0, \tilde{\xi} \Xi \tilde{\xi}^T)$$

where $\tilde{\xi} = (Q^T \otimes Q^{-1}) \xi$ *and* $\tilde{b} = \alpha \Sigma_X^{11} (G_{11} G_{11}^T - G_{11}^{:m} (G_{11}^{:m})^T) \beta^T$ *is the asymptotic bias.*

The $T^{\frac{1}{2}}$ terms dominate in the limiting behaviour of $\hat{\Pi}_k$, which is why, asymptotically, we lose nothing by overestimating the rank. However, for finite samples the case might be different. As suggested by Theorem 2, the variance in the random walk direction will increase if we unnecessarily inflate the rank. See Anderson [2002b] regarding further interpretation of the asymptotics of Π.

Remark 1. Often the cointegration matrix β is of interest and one might ask to what extent the above results are then relevant. Since β is only identifiable up to post-multiplication by an invertible matrix, one needs to impose further identifiability restrictions to be able to meaningfully discuss this. To directly utilize the above results, we take a non-standard approach. In particular, we assume (without loss of

generality) that the j'th diagonal of $L = Q^T G$ is non-zero for all $j = 1, \dots, r$ and take $\beta = L^{:r}(\text{diag}(L_{11}, \dots, L_{rr}))^{-1}$. By Assumption 3, $L^{:r}$ is identifiable up to post-multiplication by a diagonal matrix and the given β is therefore identifiable. Our estimator is then simply obtained by the same formula but replacing L with $\hat{L}$ and r with the given rank choice k. Then use the same argument as in the proof of Theorem 3 to show that the individual columns of $\hat{\beta}$ satisfy a central limit theorem. In particular, start from Lemma 3 and then apply the delta method using that eigenvectors associated with simple eigenvalues are differentiable (Theorem 8.9 in Magnus and Neudecker [2019]). When the rank is underestimated this corresponds to only taking the first $k < r$ columns of $\hat{\beta}$. Then the estimator is clearly biased but estimates the cointegration space corresponding to the k largest eigenvalues, that is, the first k columns of β. This generalizes the consistency result obtained as part of Lemma 1 of Cavaliere et al. [2012].

2.4 Estimation Under Rank Uncertainty

The above results suggest that the choice of cointegration rank is crucial. While choosing a rank that is too high still results in a consistent estimator, underestimating the rank will result in an asymptotically biased estimator. The rank is usually found using a sequential testing approach as described in Johansen [1995], which we briefly recall here. While this approach consistently estimates the true rank (at least if the critical values of the sequential tests go to infinity at an appropriate rate with increasing sample size), disregarding the uncertainty involved in rank estimation from the corresponding post-selection reduced rank estimator might be unfavourable in some cases, especially for high dimension p. We therefore suggest a weighted estimator of Π, which can be thought of as a weighted average of the estimators $\hat{\Pi}_1, ..., \hat{\Pi}_p$ with either fixed pre-specified weights or with weights inferred from the data. The post-selection estimator obtained by considering the rank-estimate as fixed is a special case where all the weight is assigned to $\hat{\Pi}_{\hat{r}}$, $\hat{r}$ being the rank-estimate.

2.4.1 Rank Selection

We start with the hypothesis $H(0)$ that the cointegration rank is 0, that is, Π vanishes so that the process is a random walk. This null-hypothesis is tested either against $H(1)$ or $H(p)$, which are the hypotheses for cointegration rank 1 and p, respectively. The latter hypothesis corresponds to Π having full rank and thus, under the current assumptions, to a stationary process. If $H(0)$ is rejected at, say, a 5% significance level, we move on to the next hypothesis $H(1)$ and, again, test it either against $H(2)$ or $H(p)$. This process is repeated until reaching an r for which $H(r)$ cannot be rejected. Assuming that Z_0 is Gaussian, we can directly compute the maximized likelihood function of each hypothesis and thus also a likelihood ratio test statistic. For testing $H(r)$ against $H(p)$ the likelihood ratio statistic, $LR(H(r)|H(p))$, is given by

$$-2 \log LR(H(r)|H(p)) = -T \sum_{i=r+1}^{p} \log(1 - \hat{\lambda}_i).$$

The likelihood ratio statistic for testing $H(r)$ against $H(r+1)$ is given by

$$-2\log LR(H(r)|H(r+1)) = -T\log(1-\hat{\lambda}_{r+1}).$$

The two test statistics, depending on whether we test against $H(p)$ or $H(r+1)$, are usually called the trace and maximum eigenvalue test statistics, respectively. The asymptotics of either can be derived from our discussions above. Assuming that the true rank is r_0, both statistics tend to infinity in probability for $r < r_0$. The null is therefore rejected when the statistic is larger than a critical value c. In practice, we therefore expect to underestimate the rank in cases where one or more of the population eigenvalues $\lambda_1, \cdots, \lambda_r$ are close to 0. Recall that the bias is determined only by the smallest $r-m$ of these eigenvalues where $m < r$ is the estimated rank and r the true rank.

Let $lr = (lr_0, lr_1, ..., lr_{p-1})$ denote a sequence of test statistics with either $lr_k = -2\log(H(k)|H(p))$ or $lr_k = -2\log(H(k)|H(k+1))$ and let $c_T = (c_{T,0}, ..., c_{T,p-1}) \in \mathbb{R}^p_+$ be a sequence of critical values. A rank estimate is then given by

$$\hat{r} = \min\left(\inf\left\{0 \leq k \leq p | lr_k \leq c_{T,k}\right\}, p\right) \tag{2.4.18}$$

where we use the convention $\inf \emptyset = \infty$. Usually $c_{T,k}$ is chosen to be the $(1-\alpha)100\%$ quantile of the asymptotic distribution of either the trace or maximum eigenvalue test-statistic for some small $\alpha \in (0,1)$. Letting α approach 0 with growing sample size at an appropriate rate ensures that $\hat{r}$ is a consistent estimator of the true rank, r_0.

2.4.2 Weighted Reduced Rank Estimator

The reduced rank estimators of Γ discussed thus far can all be considered as special cases of a general family of estimators weighting the contribution of each of the eigenvectors in (2.3.10). Indeed, for any $1 \leq k \leq p$, we can write

$$\hat{\Gamma}_k = S_{\Delta XX}\sum_{i=1}^{k}\hat{g}_i\hat{g}_i^T = S_{\Delta XX}\sum_{i=1}^{d} w_i\hat{g}_i\hat{g}_i^T \tag{2.4.19}$$

where $w_i = 1$ if $i \leq k$ and $w_i = 0$ otherwise. For a given vector of weights, $w \in [0,1]^p$, with $w_1 \leq w_2 \leq ... \leq w_p$, we refer to the estimator given by (2.4.19) as the *weighted reduced rank estimator* of Γ and write $\hat{\Gamma}_w$. It is also entirely possible to choose weights that depend on the data. Thus, the post-selection estimator, $\hat{\Gamma}_{\hat{r}}$, with $\hat{r}$ as given in (2.4.18), can be written as a weighted reduced rank estimator with weights

$$\hat{w}_i = \mathbf{1}_{\{i \leq \hat{r}\}} = \mathbf{1}_{\{lr_{i-1} > c_{T,i-1}\}}. \tag{2.4.20}$$

Furthermore, the weighted reduced rank estimator can be viewed as a weighted average of all the individual reduced rank estimators. To see this, define the additional weights $w_0 = 0$ and $w_{p+1} = 1$ and let $W_i = w_{i+1} - w_i$ for $i = 0, ..., p$. Then $W \in [0,1]^{p+1}$ with $\sum W_i = 1$ and

$$\hat{\Gamma}_w = \sum_{k=0}^{p} W_k\hat{\Gamma}_k$$

with the convention $\hat{\Gamma}_0 = 0$.

Assuming that the true rank is $1 \leq r_0 \leq p$ it is immediately clear from Section 2.3 that any weighting that does not asymptotically assign weight one to all eigenvectors, $\hat{g}_1, ..., \hat{g}_{r_0}$, will result in an asymptotically biased estimator. Conversely, if $w_i \to_p 1$ for all $i = 1, ..., r_0$, then $\hat{\Gamma}_w$ is consistent regardless of the asymptotic behaviour of the rest of the weights. Now, for $w \in [0,1]^p$, let $D = \text{diag}(w)$ and write $D_1 = (D_{i,j})_{i,j \leq r}$ and $D_2 = (D_{i,j})_{i,j>r}$. We define the following quantities:

$$\begin{aligned} b_w &= \beta^T \alpha \Sigma_X^{11} G_{11}(I_r - D_1) G_{11}^T, \\ C_{1w} &= \beta^T \alpha \Sigma_X^{11} G_{11} D_1 G_{11}^T \left(\beta^T \alpha\right)^{-1}, \\ C_{2w} &= G_{22} D_2 G_{22}^T, \\ \xi_w &= \sum_{i=1}^{r} w_i \left(\begin{pmatrix} 0_{r \times p} & P_i \end{pmatrix} \otimes \begin{pmatrix} I_p & 0_{p \times r} \end{pmatrix} + (I_r \otimes \Sigma_{\Delta X X}) \xi_i \right) \end{aligned}$$

where ξ_i is defined in (2.A.6) and $P_i = G_{11} e_i (G_{11} e_i)^T$ with e_i being the i'th unit vector, that is, $G_{11} e_i$ is the i'th column of G_{11}. The following result is a consequence of Theorems 1, 2 and 3.

Theorem 5. *Assume that $1 \leq r = \text{rank}(\Gamma) \leq p$ and $\lambda_1 > \cdots > \lambda_r$. Let $(w_T)_{T \in \mathbb{N}} \subset [0,1]^p$ be a sequence of weights. Assume that $w_{T,i} \to_p 1$ for all $i = 1, ..., r$. Then, for $T \to \infty$,*

$$\hat{\Gamma}_{w_T} \to_p \Gamma.$$

Assume furthermore that $T||(w_T - w)|| \to_p 0$ for some $w \in [0,1]^p$ and $\kappa_{ijkl} = 0$ for all $1 \leq i, j, k, l \leq p$. Let $V_w^T = (\tilde{V}_{w,11}^T, \tilde{V}_{w,21}^T)$ be a random matrix such that $\text{vec}(V_w) \sim \mathcal{N}(0, \xi_w \Xi \xi_w^T)$, where Ξ is defined in (2.A.5) and ξ_w is defined above. Then,

$$\begin{pmatrix} T^{\frac{1}{2}}(\hat{\Gamma}_{w_T}^{11} - \Gamma_{11} - b_w) & T(\hat{\Gamma}_{w_T}^{12} - \Gamma_{12}) \\ T^{\frac{1}{2}}(\hat{\Gamma}_{w_T}^{21} - \Gamma_{21}) & T(\hat{\Gamma}_{w_T}^{22} - \Gamma_{22}) \end{pmatrix} \to_w \begin{pmatrix} V_{w,11} & C_{1w} \tilde{J}_{12} B^{-1} + \tilde{J}_{22} C_{2w} \\ V_{w,21} & J_{22} C_{2w} \end{pmatrix}. \tag{2.4.21}$$

It follows from Theorem 5 that if a sequence of weights, $(w_T)_{T \in \mathbb{N}}$, is such that $T(w_{T,i} - \mathbf{1}_{\{i \leq m\}}) \to_p 0$ for some $1 \leq m \leq p$, then $\hat{\Gamma}_{w_T}$ is asymptotically equivalent to $\hat{\Gamma}_m$. In particular, for the post-selection estimator (2.4.18) with weights (2.4.20), for every $\epsilon > 0$, we have

$$\begin{aligned} \mathbb{P}\left(T|\hat{w}_i - \mathbf{1}_{\{i \leq r\}}| > \epsilon\right) &\leq \mathbb{P}\left(\hat{r} \neq r\right) \\ &= \mathbb{P}\left(\hat{r} > r\right) + \mathbb{P}\left(\hat{r} < r\right) \\ &\leq \sum_{i=0}^{r-1} \mathbb{P}\left(lr_i \leq c_{T,i}\right) + \mathbb{P}\left(lr_r > c_{T,r}\right) \end{aligned}$$

where the last term goes to 0 for $T \to \infty$ if c_T is chosen appropriately, since lr_i goes to infinity for $i < r$ and lr_r converges in distribution and is therefore bounded in probability.

Consequently, the post-selection estimator is asymptotically equivalent to $\hat{\Pi}_r$.[1] In finite samples, however, the situation can be different. It is entirely possible that lr_i is close to $c_{T,i}$ for multiple $i = 0, ..., r$, which would indicate that there is evidence for multiple ranks in the observed data. Hard threshold weights like $\hat{w}$ disregard this uncertainty and for some samples the choice of rank can be far from the true rank (see Appendix 2.D.3). It might therefore be wise to explore weights that behave more smoothly. We here give two examples of such weights both of which are based on the likelihood-ratio test statistics, lr.

Example 2.4.1. The first weight-vector we consider is motivated by the fact that large values of lr_i are strong evidence that the eigenvector $\hat{g}_{i+1}$ should be included. Indeed, as stated above, if $i < r$, then $T^{-a}lr_i$ goes to infinity for any $a \in [0, 1)$. It is similar to the weighting scheme considered in Lieb and Smeekes [2017]. For $a_1 > 0$ and $0 \leq a_2$ we define

$$\hat{w}_1(a_1, a_2) = \left(1 - e^{-a_1 T^{-a_2} lr_0}, ..., 1 - e^{-a_1 T^{-a_2} lr_{p-1}}\right). \tag{2.4.22}$$

When convenient we shall omit the arguments and simply write $\hat{w}_1$. The hyperparameters a_1 and a_2 control how sensitive the weights are to the size of $\hat{\lambda}$. If $1 > a_2 > 0$, then $T^{-a_2} lr_i \to_p \infty$ for $i < r$ and $T^{-a_2} lr_i \to_p 0$ for $i \geq r$ which implies that $T(\hat{w}_1 - \mathbf{1}_{\{\cdot \leq r\}}) \to_p 0$ for $T \to \infty$, i.e., $\hat{\Pi}_{\hat{w}_1}$ is asymptotically equivalent to $\hat{\Pi}_r$. ♠

Example 2.4.2. The second example is a soft threshold version of the categorical $\hat{w}$. We simply replace the indicator function in $\hat{w}_i = \mathbf{1}\{lr_{i-1} > c_{T,i}\}$ with a sigmoid function. Specifically, let $\tau : \mathbb{R} \to [0, 1]$ be a sigmoid function, i.e., monotone and differentiable with $\tau(0) = 0.5$, $\tau(-\infty) = 0$, and $\tau(\infty) = 1$, and define for $a > 0$ and $c = (c_0, ..., c_{p-1})$ the weights

$$\hat{w}_2(a, c) = \left(\tau\left(a(lr_0 - c_0)\right), ..., \tau\left(a(lr_{p-1} - c_{p-1})\right)\right). \tag{2.4.23}$$

Similar to above, we will sometimes omit the arguments for notational simplicity. In most applications it would make sense to just choose c_i to be the $(1 - \alpha)100\%$ quantile for the asymptotic distribution of the test statistic lr_i for some prespecified signifcance level $\alpha \in (0, 1)$ in which case we shall write $\hat{w}_2(a, \alpha)$. The hyperparameter a controls the gradient of the sigmoid function with higher values resulting in a sharper separation. One can choose a, c dependent on T such that $a(lr_i - c_i) \to_p \infty$ if $i < r$ and $a(lr_i - c_i) \to_p -\infty$ if $i \geq r$ in which case $\hat{\Pi}_{\hat{w}_2}$ is also asymptotically equivalent to $\hat{\Pi}_r$. This weight works well for moderate dimensions. Indeed, for very high p, choosing the appropriate vector $c \in \mathbb{R}^p$ becomes prohibitive. ♠

Remark 2. For many choices of weights (as in the examples above) one needs to specify a set of hyperparameters. This might be hard in practice and is perhaps the main drawback of the general weighted reduced rank estimators. However, in many cases it

[1] This is only true in a pointwise sense. Indeed, the situation is very different if one considers sequences of parameters Π_T with eigenvalues getting arbitrarily close to 0. See, for example, the discussion in Elliott [1998] or the simulations in the following section. For uniform asymptotic inference in this setting, see Holberg and Ditlevsen [2024b].

suffices to chose a list of candidate parameters and then pick the best based on cross-validation. For the weights given in Example 2.4.1 it often suffices to fix $a_1 = 1$ and then simply pick a_2 so as to minimize, for example, the mean squared prediction error over 10-fold cross-validation. Finally, in high dimensions, one might need to resort to cross-validation to estimate the rank anyway since the usual sequential testing approach outlined above quickly becomes infeasible.

2.5 Simulation study

In this section we perform two sets of simulation studies. First we compare different weighted reduced rank estimators across a range of parameters. The second set of simulation experiments compares the empirical large-sample distribution of our estimators with the asymptotic distributions derived in Section 2.3. This will not only confirm our results, but also give an idea of how the distributions in (2.3.14), (2.3.15), and (2.3.17) behave which is useful especially for the last case because of its complicated nature. Details about the different simulation setups are given in Appendix 2.D.

2.5.1 Comparison of weighted reduced rank estimators

We compare different weighted reduced rank estimators for a handful of configurations. We compare 4 different types of weights. The first type is given by $w_f(k) \in [0,1]^p$ with $(w_f(k))_i = 1$ for $i \leq k$ and 0 otherwise. w_f does not depend on the observed data but simply chooses a fixed number of eigenvectors to include. It thus corresponds to the simple reduced rank estimators of fixed rank, i.e., $\hat{\Gamma}_{w_f(k)} = \hat{\Gamma}_k$. The second type is the post-selection weight $\hat{w} = w_f(\hat{r})$ in (2.4.20). The last two types are $\hat{w}_1$ and $\hat{w}_2$ in eqs. (2.4.22) and (2.4.23) for different values of hyperparameters. We consider $w_2(a, \alpha)$ as a function of the first parameter a and a significance level $\alpha \in (0,1)$ as explained in Example 2.4.2. We used the error function as the sigmoid function, i.e.,

$$\tau(x) = \frac{\operatorname{erf}(x) + 1}{2}, \quad \operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt.$$

We are interested in settings where there might be uncertainty regarding the choice of rank. This simulation experiment therefore considers sequences of parameters $\Gamma_c \in \mathbb{R}^{p \times p}$ for which a third of the eigenvalues are stationary (fixed at $-3/2$), a third is exactly 0, and the last third is given by $-c/T$ for varying $c \geq 0$. Throughout we fix $T = 100$. For $p = 3, 6, 9$ and $c \in [0, 30]$ we then compare all the estimators based on the mean squared error (MSE) and the mean squared prediction error (MSPE), which, for an estimator $\hat{\Gamma}$, is given by $T\mathbb{E}||\Delta X_{T+1} - \hat{\Gamma} X_T||^2$. For a detailed description, see Appendix 2.D.1. The results for the MSPE given in Figure 2.5.1 are based 4 million simulations. The lines in the figure are smooth versions of the actual results reported in Appendix 2.D. The results for the MSE in Figure 2.5.2 are based on 50 thousand simulations and there was no need to do any smoothing. For a decomposition of the MSE in terms of bias and variance we refer to Figure 2.D.4 and 2.D.5 in Appendix 2.D.

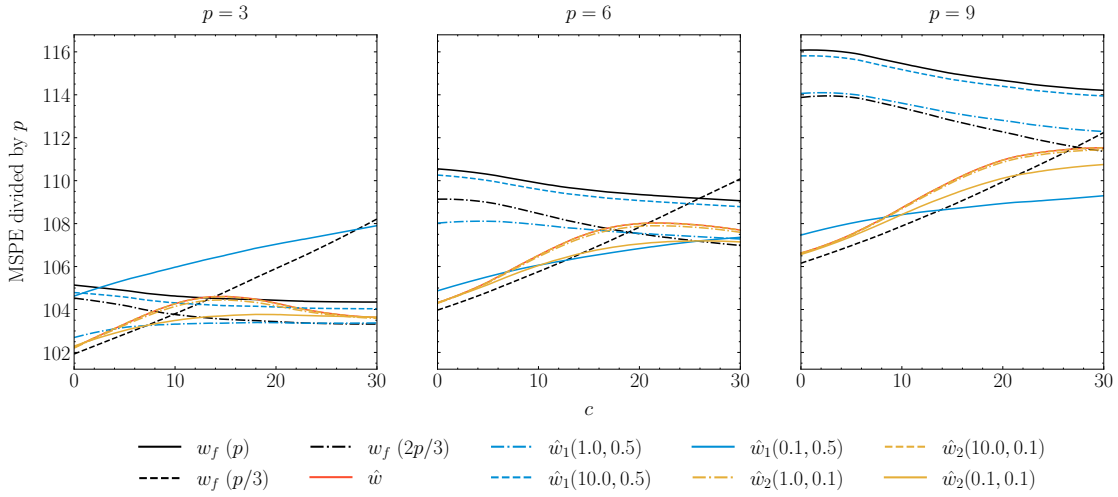

Figure 2.5.1: Mean square prediction error (MSPE) of different weighted reduced rank estimators for varying dimensions and $c \in [0, 30]$ where the underlying autoregressive matrix, Γ_c, has a third of its eigenvalues set to $-c/T$, a third set to 0 and a third set to $-3/2$. Sample size is fixed at $T = 100$. The lines have been smoothed out for better comprehension. See Figure 2.D.1 for the true graphs.

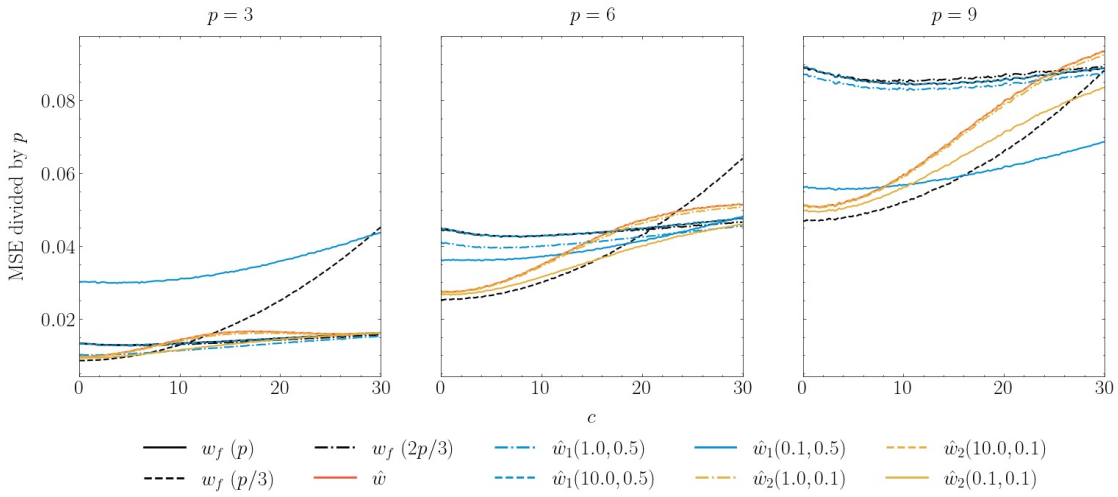

Figure 2.5.2: Mean square error (MSE) of different weighted reduced rank estimators for varying dimensions and $c \in [0, 30]$ where the underlying autoregressive matrix, Γ_c, has a third of its eigenvalues set to $-c/T$, a third set to 0 and a third set to $-3/2$. Sample size is fixed at $T = 100$. See Figure 2.D.4 and 2.D.5 for bias-variance decomposition.

For the parameters considered here, the cointegration rank will always be equal to two-thirds of the dimension p, although, for c close to 0, it will be practically $p/3$. Thus, the least squares estimator is really overparameterized which results in a higher variance (see Figure 2.D.5) and thus also a higher MSE in almost all cases. This is in line with the asymptotic theory developed in Section 2.3. Interestingly, the reduced rank estimator with $k = 2p/3$ has a very similar performance in terms of MSE in all cases. Much of this can probably be attributed to the uncertainty in the ordering of eigenvalues for finite samples. For c close to 0, the reduced rank estimator with $k = p/3$ tends to perform the best for all dimensions considered here. This advantage quickly disappears, though, when c increases as a result of the bias that is introduced (see also Figure 2.D.4). As we have tried to emphasize, in practice one will not know the true rank so a more realistic estimator of Γ to look at in this case is the post-selection estimator also taking rank estimation into account. In all cases we see that the post-selection estimator is outperformed by the weighted reduced rank estimators using the $\hat{w}_2$ weights regardless of the hyperparameters considered. For $\hat{w}_1$ it seems that the choice of hyperparameters plays a much larger role. While $\hat{w}_1(0.1, 0.5)$ is clearly the worst for $p = 3$, it actually outperforms all others for $p = 9$ and for c larger than, say, 17.

In terms of MSPE we see similar results. The least squares estimator has the highest MSPE in almost all cases. For c small enough, the reduced rank estimator of rank $p/3$ outperforms the reduced rank estimator of rank $2p/3$ (the dashed and dash-dotted black lines). Thus, choosing a rank smaller than the true rank is beneficial if the discarded eigenvectors have eigenvalues close enough to 0. At some point, however, c is too large and the bias induced by discarding these eigenvector will grow correspondingly at which point the reduced rank estimator of rank $2p/3$ is preferable (around $c = 14, 19, 27$ from left to right). All the data-dependent weights are attempting to detect this point and act accordingly. For the estimators based on $\hat{w}_1$ it seems as though the MSPE is shifted depending on the dimension. In higher dimension $\hat{w}_1(0.1, 0.5)$ is a good choice while $\hat{w}_1(1, 0.5)$ outperforms the other estimators most of the time for $d = 3$. Similarly, smoothing the rank-selection weights increases the predictive performance of the estimator. Indeed, the weighted reduced rank estimator with weights $\hat{w}_2(0.1, 0.1)$ clearly outperforms the post-selection estimator in all cases. In Figures 2.D.2 and 2.D.3 in Appendix 2.D.1 we have plotted the mean and standard deviation of all the weights across all simulations for $p = 3$ which potentially explains a lot of the differences in performance. Similar behaviour holds for $p = 6$ and $p = 9$.

2.5.2 Comparison of asymptotic and empirical distributions

We consider estimators of Γ, comparing each block separately. The dimension is $p = 4$, the true rank is chosen to be $r = 2$ and we consider the estimators $\hat{\Gamma}_1$, $\hat{\Gamma}_2$, and $\hat{\Gamma}_4$ corresponding to the three cases of underestimated, correct and overestimated rank. We let Z_t be i.i.d. normal with $Z_t \sim \mathcal{N}(0, \Sigma_Z)$. We generate $\alpha, \beta \in \mathbb{R}^{4\times 2}$ and $\Sigma_Z \in \mathbb{R}^{4\times 4}$ such that Assumptions 1, 2, and 3 are fulfilled. For explicit details on the simulation setup, see Appendix 2.D.

In Fig. 2.5.3 we compare the empirical distributions and the asymptotic distributions

of the three estimators. That is, we compare the distributions on the left-hand sides to the right-hand sides of (2.3.14), (2.3.15), and (2.3.17). For the estimators under underestimated rank, we also subtracted the bias in eq. (2.3.16), which is why it appears to be centered. Observe that the estimators for the true rank (red lines) are not visible in most plots because they overlap with the other lines. This agrees with the theory. For the two leftmost columns, the distribution of $\hat{\Gamma}_2^{\cdot 1}$ coincides with the distribution of $\hat{\Gamma}_4^{\cdot 1}$. For the bottom-right block, $\hat{\Gamma}_2^{22}$ and $\hat{\Gamma}_1^{22}$ are both singular around 0 which is why the empirical distributions are highly concentrated compared to the empirical distribution of $\hat{\Gamma}_4$.

For all three estimators, the large-sample empirical distribution is close to the asymptotic distribution which confirms our theoretical findings. Furthermore, as we hypothesised, the variance of $\hat{\Gamma}_1$ is decreased in all blocks except the top-left block. The decrease seems to be most visible in the bottom-left block. This is surprisingly not the case when we compare $\hat{\Gamma}_1^{11}$ with $\hat{\Gamma}_2^{11}$. It looks like the former has a higher variance in some of the elements. In other simulations the results were also ambiguous making any quantitative judgements hard. It should be noted, however, that in Fig. 2.5.3 the distribution of each element of the estimated matrix is plotted separately, i.e., we do not consider the covariance structure between different elements of the matrix.

2.6 Prediction of EEG Signals

We apply our weighted reduced rank estimators to EEG recordings obtained from an experiment in which two participants were presented with a visual stimulus on a computer screen. Each participant was first shown a cross on the screen on which to focus for a random fixation period between 1.5 and 2.5 seconds. Then, two figures would briefly appear on the screen and the participant should indicate which stimulus had been shown. For more information on the exact setup, see Levakova et al. [2022]. Here, we analyze the trials from participant 1 which, after data clean-up, amount to 609 trials in total with a sampling rate of 256 Hz. For each trial we only consider the period one second prior to the onset of the visual stimulation. The psychological hypothesis is that the brain state at stimulus onset is predictive of cognitive performance, and this short pre-stimulus period is therefore of special interest. After data preprocessing and clean-up, each EEG signal consists of 59 channels. The resulting data set has 609 observations of 59-dimensional time series of sample size 257. We represent the data by $X_t^i \in \mathbb{R}^{59}$ where $i = 1, \dots, 609$ and $t = 0, 1, \dots, 256$. See Fig. 2.6.4 for a sample of X^i.

We analyze the predictive capabilities of the weighted reduced rank estimators for two classes of weights on the given data. We consider the discrete weights given by $w_f(k)$, $k = 0, \dots, 59$ as well as the smooth weights $\hat{w}_1(1, a_2)$ for $a_2 = k/25$, $k = 0, \dots, 49$. The high dimension of the data makes any classical methods of rank selection as well as methods based on bootstrap prohibitive. Similarly, the weights given in Example 2.4.2 are not well suited for problems in higher dimension due to the need to select the thresholds $c \in \mathbb{R}^d$. The methods proposed so far in this paper are in the setting of a single observation of a long time series and under the assumption of zero drift. They are,

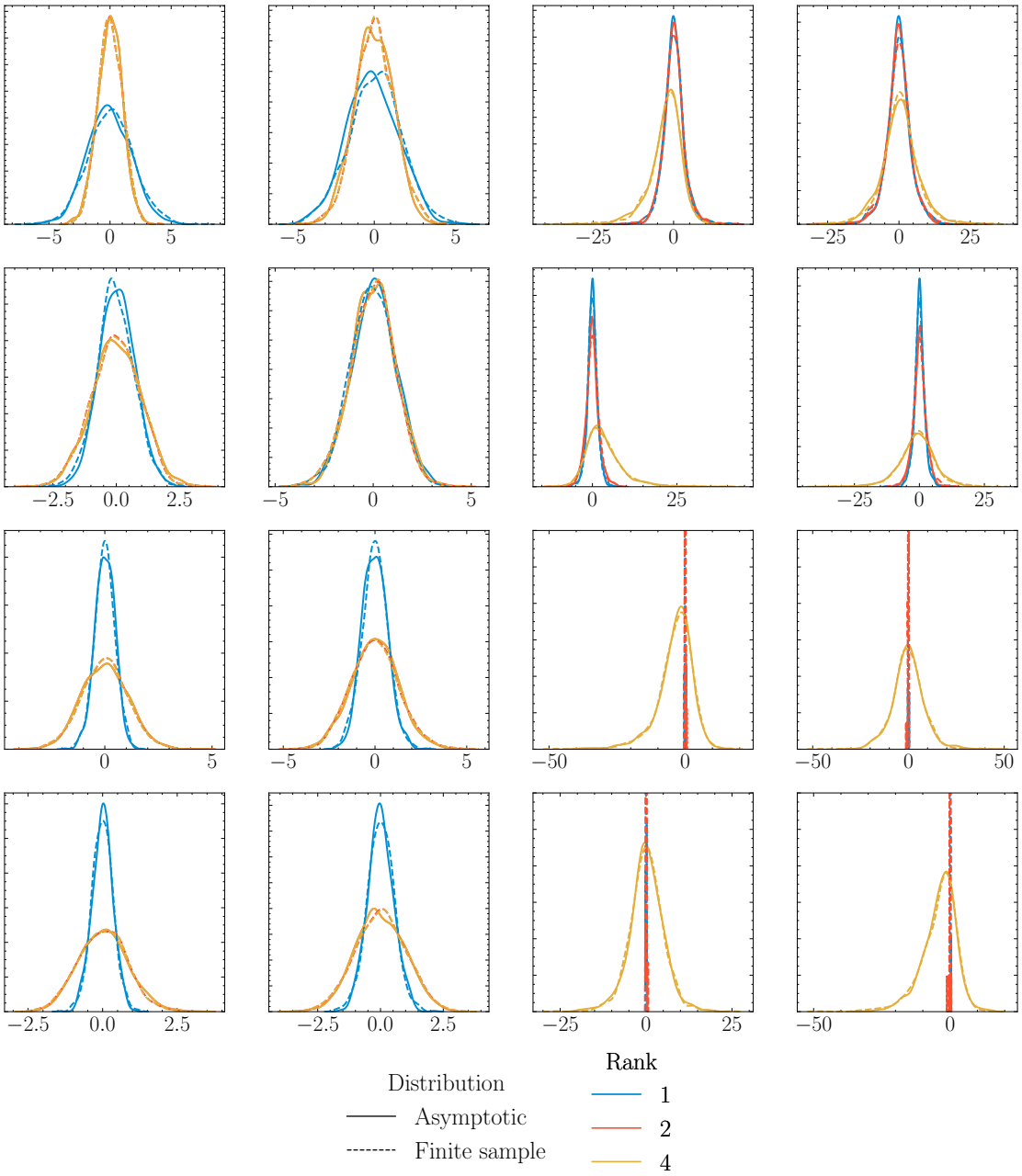

Figure 2.5.3: Asymptotic and empirical distributions of $\hat{\Gamma}_k - \Gamma$ for different choices of k. The dimension is $p = 4$, the true rank is $r = 2$ and the three estimators are $\hat{\Gamma}_1$, $\hat{\Gamma}_2$, and $\hat{\Gamma}_4$. For each estimator, the dotted line is the empirical distribution for $T = 5000$ and over 1000 simulations. The i, j'th plot corresponds to the distribution of the i, j'th element of $\hat{\Gamma}_k - \Gamma$. We centered $\hat{\Gamma}_1 - \Gamma$ by subtracting the bias given in the right hand side of (2.3.16) in Section 2.3. Note different scales in individual plots.

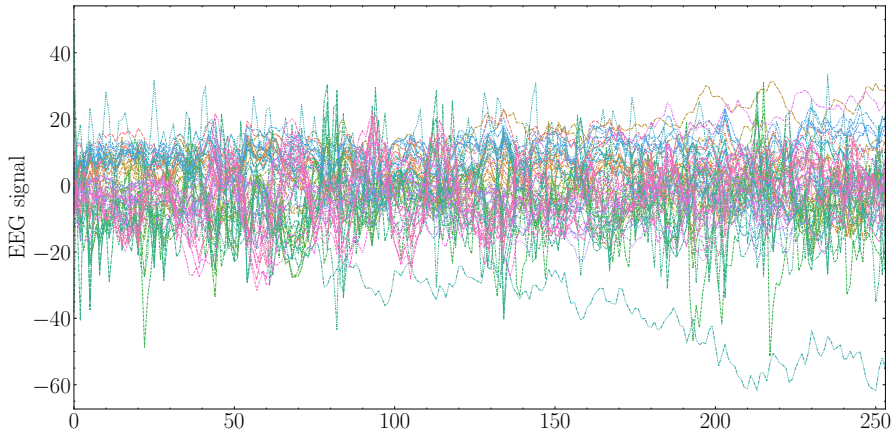

Figure 2.6.4: Plot of a sample of 59 EEG channels from participant 1 ranging over a second prior to stimulation onset and sampled at 256 Hz.

however, straightforwardly adapted to work in settings with multiple i.i.d. observations of the same time series and to allow for the inclusion of a constant drift [Levakova et al., 2022, Section 2].

In Fig. 2.6.5 we record the performance of each $w_f(k)$ and $\hat{w}_1(1, a_2)$ in terms of MSPE. For a test/train split $I_{train}, I_{test} \subset \{1, \dots, 609\}$ with $I_{train} \cap I_{test} = \emptyset$, the model was fitted on the train set $(X^i)_{i \in I_{train}}$ and the MSPE calculated on the test set $(X^i)_{i \in I_{test}}$. We compare the estimators for three different sample sizes of the training data. A train size of q and test size of p means that $|I_{train}| = \lfloor q609 \rfloor$ and $|I_{test}| = \lfloor p609 \rfloor$. Throughout we fix the test size at 0.1, i.e., 10% of the observations are used to compute the MSPE. The results reported in Fig. 2.6.5 are averaged across 40 random test/train splits of the data for train sizes 0.1, 0.2 and 0.3.

Evidently, for both choices of weights, the fixed rank and the smoothed weights, the hyperparameter, r or a_2, strongly affects the performance of the corresponding estimator. A similar pattern emerges in the left and right panel of Fig. 2.6.5. At certain thresholds the predictive capabilities plateau around the same level, namely, for $a_2 \leq 0.9$ and $r \geq 25$. One way to interpret this is that, after a while, increasing the rank of the estimator does not yield better results, i.e., we lose nothing by using a lower rank representation of the underlying dynamics. Similarly, for $a_2 \geq 1.5$, $\hat{w}_1(1, a_2)$ is practically 0 so that the MSPE in the left panel plateaus at the same level as the MSPE for $w_f(0)$. Interestingly, whereas the MSPE in the right panel seems to be almost monotone in the hyperparameter, this is not the case in the left panel. Especially for the smallest train size, the MSPE of $\hat{w}_1$ dips well below the lowest level achieved by w_f. Thus, for small sample sizes, the new estimator with smooth weights performs better. In practice, we do not know the optimal choice of a_2 or k, but this can be partly resolved by cross-validation. Fig. 2.6.6 depicts the distribution of the MSPE corresponding to the reduced rank estimators with weights $w_f(\hat{k}_{cv})$ and $\hat{w}_1(1, \hat{a}_{2,cv})$ where $\hat{k}_{cv}$ and $\hat{a}_{2,cv}$ were chosen to yield the lowest MSPE based on cross-validation on the training data with 10 folds.

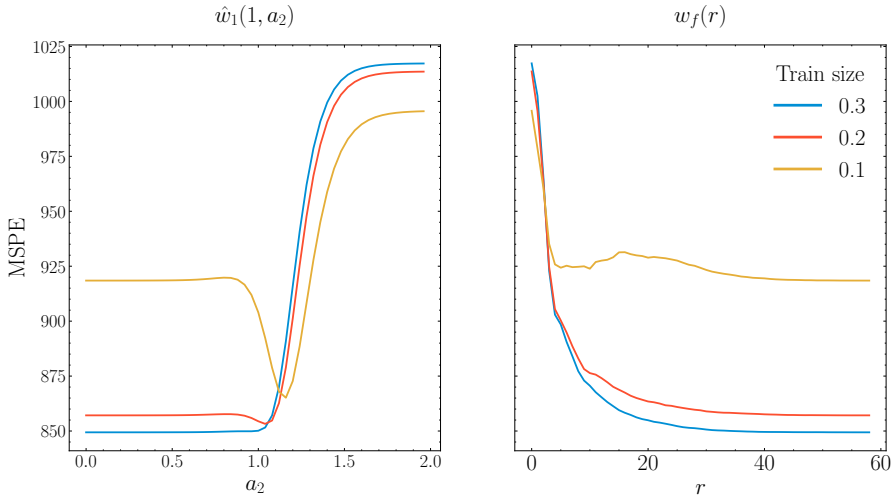

Figure 2.6.5: Average mean square prediction error (MSPE) for chosen reduced rank estimators over 40 random test/train splits of the data for three different choices of train size. In each case the test size is fixed at 0.1.

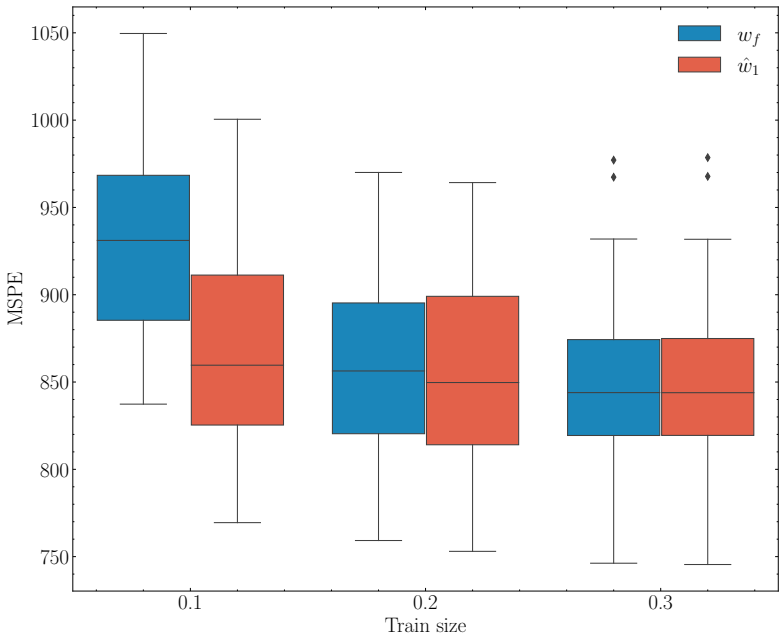

Figure 2.6.6: Distribution of the mean square prediction error (MSPE) corresponding to the estimators with weights $w_f(\hat{k}_{cv})$ and $\hat{w}_1(1, \hat{a}_{2,cv})$. For each split, the hyperparameters $\hat{k}_{cv}$ and $\hat{a}_{2,cv}$ were chosen by cross-validation on the training data.

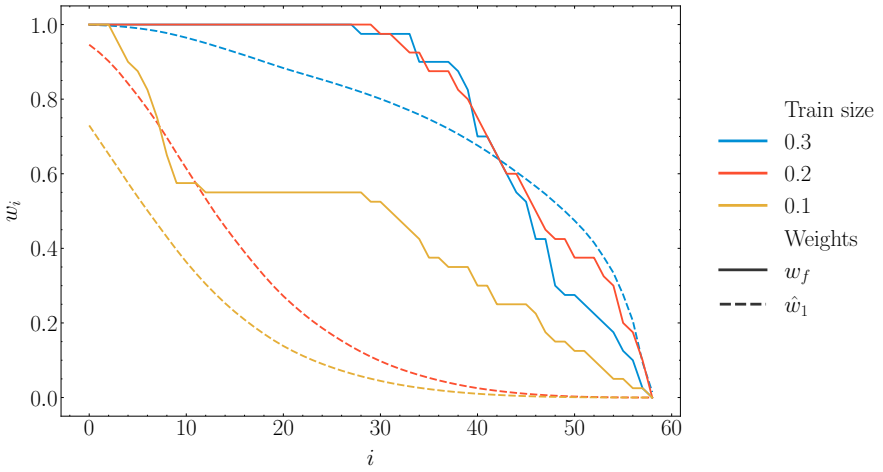

Figure 2.6.7: Average values of the weights $w_f(\hat{k}_{cv})_i$ for the classical fixed rank estimator (i.e., weights are 1 for $i \leq \hat{k}_{cv}$ and 0 otherwise) and $\hat{w}_1(1, \hat{a}_{2,cv})_i$ for the new proposed weighted rank estimator, $i = 1, \ldots, 59$, across 40 test/train splits for different choices of train size. For each split, the hyperparameters $\hat{k}_{cv}$ and $\hat{a}_{2,cv}$ were chosen by cross-validation on the training data.

For train sizes 0.2 and 0.3 the two estimators seem to do equally well. This in line with the results in Fig. 2.6.5. However, the situation changes for the smallest train size where the smooth weights clearly outperform the discrete weights. In Fig. 2.6.7 we can see that the weights behave differently. The smooth weights tend rather quickly to 0 for larger ranks as the sample size decreases, but the discrete weights are slower to react. In particular, for the smallest sample the chosen rank varies a lot based on the particular data split. This shows that smoothing the weights is beneficial in settings with large rank uncertainty (in this case because of the high dimension and the small sample size).

2.7 Conclusion

We have characterised the asymptotic distribution of all reduced rank estimators of the Π-matrix in a VECM as given by (2.1.1) assuming the cointegration rank, r, and the dimension, p, are held fixed and the sample size $T \to \infty$. Previously, only the asymptotic distribution of the reduced rank estimator with true rank has been studied. We showed that similar results hold if the rank is respectively overestimated or underestimated. In the first case, the Q-transformed estimator is still consistent albeit at the cost of an increased variance. In particular, the bottom-right block of (2.3.15) no longer converges in probability to zero. This is to be expected since we are effectively including more parameters than necessary. In the original coordinates this increased variance appears in the random walk direction and is not asymptotically relevant on the $\sqrt{T}$ scale (see Theorem 4). In the second case, the estimator is asymptotically biased and the size of the bias is determined by how much the rank is underestimated. Simulation studies

confirmed the theoretical findings.

We have introduced a new class of estimators that outperform the classical estimators in settings where certain eigenvalues are close to zero. By choosing appropriate weights that take rank uncertainty into account, the weighted reduced rank estimators have several benefits. They are transparent regarding rank evidence, they have smaller mean square prediction error and the resulting estimators have less variance when compared to the post-selection estimator. When applied to high-dimensional EEG data we find that the weighted reduced rank estimators perform favourably for small sample sizes. In order to shine some light on this behaviour, an interesting avenue for future research would be to study the asymptotics of the different quantities in a high-dimensional setup, that is, asymptotic regimes where the dimension p diverges with the sample size.

Appendix

2.A. Proofs

Proof of Theorem 1. For ease of notation we write $A := S_{X\Delta X}(S_{\Delta X\Delta X})^{-1}S_{\Delta XX}$. From Lemma 1 we find that $T^{-\frac{1}{2}}A_{12}$, $T^{-\frac{1}{2}}A_{22}$, $T^{-\frac{1}{2}}S^{21}_{XX}$, and $T^{-\frac{1}{2}}S^{12}_{XX}$ are $o_P(1)$. Since S^{11}_{XX} and $T^{-1}S^{22}_{XX}$ are $O_P(1)$, from (2.3.11) we get that $(\hat{G}_{11}, \hat{G}_{12})$ and $T^{\frac{1}{2}}(\hat{G}_{21}, \hat{G}_{22})$ are bounded in probability for all $j = 1, ..., p$. It follows that $A_{12}\hat{G}_{21}$, $A_{22}\hat{G}_{21}$, and $S^{12}_{XX}\hat{G}_{21}$ are $o_P(1)$. Finally, note that $\hat{\Lambda}_{11}$ defined in (2.3.10) converges in probability to matrix Λ_{11} and $\hat{\lambda}_i = O_P(T^{-1})$ for $i = r+1, ..., p$ (see e.g. Johansen [1988]).

Writing (2.3.10) in block matrix notation we then have for $\hat{G}^{:r}$,

$$A_{11}\hat{G}_{11} + o_P(1) = S^{11}_{XX}\hat{G}_{11}\hat{\Lambda}_{11} + o_P(1) \tag{2.A.1}$$

$$A_{21}\hat{G}_{11} + o_P(1) = (S^{21}_{XX}\hat{G}_{11} + T^{-1}S^{22}_{XX}T\hat{G}_{21})\hat{\Lambda}_{11}. \tag{2.A.2}$$

With $H_1 = \left(\frac{1}{T}S^{22}_{XX}\right)^{-1}\left(A_{21}(A_{11})^{-1}S^{11}_{XX} - S^{21}_{XX}\right)$ we compute $T\hat{G}_{21} = H_1\hat{G}_{11} + o_P(1)$ which, in particular, implies that $T\hat{G}_{21}$ is bounded in probability and therefore $\hat{G}^T_{21}S^{22}_{XX}\hat{G}_{21}$, $\hat{G}^T_{21}S^{21}_{XX}\hat{G}_{11}$, and $\hat{G}^T_{11}S^{12}_{XX}\hat{G}_{21}$ are all $o_P(T^{-\frac{1}{2}})$. Applying (2.3.11) then yields

$$\hat{G}^T_{11}S^{11}_{XX}\hat{G}_{11} = I_r + o_P(T^{-\frac{1}{2}}),$$

i.e., $\hat{G}_{11}\hat{G}^T_{11} = (S^{11}_{XX})^{-1} + o_P(T^{-\frac{1}{2}})$. Then we simply compute the estimator $S_{\Delta XX}\hat{G}^{:r}(\hat{G}^{:r})^T$ in (2.3.12) using the block expressions derived above. This gives us

$$\begin{aligned}
\hat{\Gamma}^{11}_r &= S^{11}_{\Delta XX}\hat{G}_{11}\hat{G}^T_{11} + o_P(T^{-\frac{1}{2}}) = \beta^T\alpha + S^{11}_{UX}(S^{11}_{XX})^{-1} + o_P(T^{-\frac{1}{2}}) \\
\hat{\Gamma}^{21}_r &= S^{21}_{\Delta XX}\hat{G}_{11}\hat{G}^T_{11} + o_P(T^{-\frac{1}{2}}) = S^{21}_{UX}(S^{11}_{XX})^{-1} + o_P(T^{-\frac{1}{2}}) \\
\hat{\Gamma}^{12}_r &= T^{-1}S^{11}_{\Delta XX}\hat{G}_{11}\hat{G}^T_{11}H^T_1 + o_P(T^{-1}) = T^{-1}\beta^T\alpha H^T_1 + o_P(T^{-1}) \\
\hat{\Gamma}^{22}_r &= o_P(T^{-1})
\end{aligned}$$

Appealing to Lemma 1 we see that $H^T_1 \to_w (\beta^T\alpha)^{-1}(J_{12} - \Sigma^{12}_W(\Sigma^{22}_W)^{-1}J_{22})B^{-1}$ jointly with (2.2.5), (2.2.6), and (2.2.7). The result of Theorem 1 is then easily derived from the above expressions. □

Proof of Theorem 2. The main ideas of this proof are similar to those of Theorem 1 and we shall proceed in the same manner. Slightly abusing the notation used so far we let $\hat{G}_{\cdot 2} = (\hat{G}^T_{12}, \hat{G}^T_{22})^T$. Equation (2.3.10) translates to

$$A\hat{G}_{\cdot 2} = S_{XX}\hat{G}_{\cdot 2}\hat{\Lambda}^{:m:m}_{22} \tag{2.A.3}$$

where $\hat{\Lambda}^{:m:m}_{22} = \mathrm{diag}(\hat{\lambda}_{r+1}, ..., \hat{\lambda}_{r+m})$. Recall that $\hat{\lambda}_i = O_P(T^{-1})$ for $i = r+1, ..., p$ so that $\hat{\Lambda}_{22} = O_P(T^{-1})$. Now since $\hat{G}_{\cdot 2} = o_P(T^{-\frac{1}{4}})$ (see the comments made at the start of Section 2.3.2) and $S^{11}_{XX}, S^{12}_{XX} = O_p(1)$, it follows that $(S^{11}_{XX}, S^{12}_{XX})\hat{G}_{\cdot 2}\hat{\Lambda}^{:m:m}_{22} = o_P(T^{-1})$. In block matrix notation the top part of eq. (2.A.3) simplifies to

$$A_{11}\hat{G}^{:m}_{12} + A_{12}\hat{G}^{:m}_{22} = o_P(T^{-1})$$

which, with $H_2 = -(A_{11})^{-1}A_{12}$, can be rewritten as $\hat{G}_{12}^{:m} = H_2\hat{G}_{22}^{:m} + o_P(T^{-1})$. Substituting this expression into the bottom part of eq. (2.A.3) and multiplying by $T^{\frac{1}{2}}$ gives

$$(A_{22} - A_{21}(A_{11})^{-1}A_{12})T^{\frac{1}{2}}\hat{G}_{22}^{:m} = S_{XX}^{22}T^{\frac{1}{2}}\hat{G}_{22}^{:m}\hat{\Lambda}_{22}^{:m:m} + o_P(T^{-\frac{1}{2}})$$

By the Davis-Kahan Theorem (see e.g. Theorem 4 in Yu et al. [2015]) there exists a random matrix L solving

$$(A_{22} - A_{21}(A_{11})^{-1}A_{12})L = T^{-1}S_{XX}^{22}LT\hat{\Lambda}_{22}^{:m:m}, \quad L^TT^{-1}S_{XX}^{22}L = I_n,$$

and such that $T^{\frac{1}{2}}\hat{G}_{22}^{:m} = L + o_P(T^{-\frac{1}{2}})$. We shall find the asymptotics of LL^T and then finish the proof by arguing that L is sufficiently close to $T^{\frac{1}{2}}\hat{G}_{22}^{:m}$. Using Lemma 1 we compute

$$\begin{aligned} H_2 &\to_w (\beta^T\alpha\Sigma_X^{11})^{-1}(\Sigma_U^{12} + \Sigma_U^{12}(\Sigma_U^{22})^{-1}J_{22}), \\ (A_{22} - A_{21}(A_{11})^{-1}A_{12}) &\to_w J_{22}^T(\Sigma_U^{22})^{-1}J_{22} \end{aligned}$$

jointly with (2.2.5), (2.2.6), and (2.2.7). With probability 1 the generalized eigenvalues on the diagonal of Λ_{22} are all distinct. Lemma 2 in Appendix 2.C along with the continuous mapping theorem then gives $LL^T \to_w G_{22}^{:m}(G_{22}^{:m})^T$ jointly with H_2 and the expressions in Lemma 1. We now have all the tools needed to evaluate $\hat{\Gamma}_{r+m}$ starting with the expression

$$\hat{\Gamma}_{r+m} = S_{\Delta XX}\begin{pmatrix}\hat{G}_{\cdot 1} & \hat{G}_{\cdot 2}\end{pmatrix}\begin{pmatrix}\hat{G}_{\cdot 1}^T \\ \hat{G}_{\cdot 2}^T\end{pmatrix} = S_{\Delta XX}(\hat{G}_{\cdot 1}\hat{G}_{\cdot 1}^T + \hat{G}_{\cdot 2}\hat{G}_{\cdot 2}^T).$$

As mentioned above $\hat{G}_{\cdot 2} = o_P(T^{-\frac{1}{4}})$ which implies that $\hat{G}_{12}^{:m}(\hat{G}_{12}^{:m})^T$ and $\hat{G}_{22}^{:m}(\hat{G}_{12}^{:m})^T$ are $o_P(T^{-\frac{1}{2}})$ and so we immediately get $\hat{\Gamma}_{r+m}^{\cdot 1} = \hat{\Gamma}_r^{\cdot 1} + o_P(T^{-\frac{1}{2}})$. For the remaining two blocks write

$$\begin{aligned} \begin{pmatrix}\hat{\Gamma}_{r+m}^{12} \\ \hat{\Gamma}_{r+m}^{22}\end{pmatrix} - \begin{pmatrix}\hat{\Gamma}_r^{12} \\ \hat{\Gamma}_r^{22}\end{pmatrix} &= \begin{pmatrix}S_{\Delta XX}^{11}\hat{G}_{12}^{:m}(\hat{G}_{22}^{:m})^T + S_{\Delta XX}^{12}\hat{G}_{22}^{:m}(\hat{G}_{22}^{:m})^T \\ S_{\Delta XX}^{21}\hat{G}_{12}^{:m}(\hat{G}_{22}^{:m})^T + S_{\Delta XX}^{22}\hat{G}_{22}^{:m}(\hat{G}_{22}^{:m})^T\end{pmatrix} \\ &= \begin{pmatrix}T^{-1}\beta^T\alpha S_{XX}^{11}H_2LL^T + T^{-1}S_{\Delta XX}^{12}LL^T \\ T^{-1}S_{\Delta XX}^{22}LL^T\end{pmatrix} + o_P(T^{-1}) \end{aligned}$$

and the result follows from Theorem 1 and Lemma 1 in combination with the limits derived above for H_2 and LL^T. □

Note that the reasoning used to determine the limit of LL^T can also be applied to $\hat{\Lambda}_{22}^{:m:m}$. In particular, Lemma 1 in the Appendix shows that $T\hat{\Lambda}_{22}^{:m:m}$ converges in distribution to $\Lambda_{22}^{:m:m}$. There is nothing special about our choice of m here and in particular $T\hat{\Lambda}_{22} \to_w \Lambda_{22}$. It is seen that

$$\begin{aligned} &|J_{22}^T(\Sigma_U^{22})^{-1}J_{22} - B\lambda| = \\ &\qquad |(\Sigma_U^{22})^{\frac{1}{2}}||(\Sigma_U^{22})^{-\frac{1}{2}}J_{22}^T(\Sigma_U^{22})^{-1}J_{22}(\Sigma_U^{22})^{-\frac{1}{2}} - (\Sigma_U^{22})^{-\frac{1}{2}}B(\Sigma_U^{22})^{-\frac{1}{2}}\lambda||(\Sigma_U^{22})^{\frac{1}{2}}|. \end{aligned}$$

Recalling the definition of J_{22} and B in Lemma 1 we get that the diagonal of Λ_{22} is, in fact, equal to the ordered solutions of

$$\left| \left(\int_0^1 W_{2s}\, dW_{2s}^T \right) \left(\int_0^1 W_{2s}\, dW_{2s}^T \right)^T - \lambda \int_0^1 W_{2s} W_{2s}^T\, ds \right| = 0$$

where W_{2s} are the last n components of the standard Brownian motion W_s. Analogously, we see from Lemma 1 and the proof of Theorem 1 that $(\hat{\lambda}_1, ..., \hat{\lambda}_r)$ are asymptotically equivalent to the ordered solutions of $|A_{11} - S_{XX}^{11} r|$. In other words, $\hat{\Lambda}_{11}$ converges in probability to Λ_{11} whose diagonal are the ordered solutions of

$$|\Sigma_X^{11} \alpha^T \beta (\Sigma_{\Delta X}^{-1})_{11} \beta^T \alpha \Sigma_X^{11} - \Sigma_X^{11} \lambda| = 0. \tag{2.A.4}$$

We have thus determined the asymptotics of $\hat{\Lambda}$ as well. This is a well known result in the cointegration literature from which one can derive the asymptotic distribution of the so-called trace test statistic, which tests the hypothesis that the cointegration rank is at most $k < p$ [Johansen, 1988].

The asymptotics are a little more involved in the case where the true rank is underestimated. Before the proof, we first show some intermediate lemmas. To study the limiting distribution of $T^{\frac{1}{2}}(\hat{\Gamma}_m^{11} - \Gamma_{11} - b_m)$ we follow the strategy of Izenman [1975] which forces us to set up more notation and introduce some ideas from matrix differential calculus. We use the notation from Magnus and Neudecker [2019]. For a matrix valued function $\Phi : \mathbb{R}^{m \times n} \to \mathbb{R}^{k \times l}$ we let $d\Phi$ denote its differential. Similarly, we define the derivative of $\Phi(A)$ with respect to A as the derivative of the vectorization of $\Phi(A)$ with respect to the vectorization of A:

$$D\Phi = \frac{\partial \text{vec}\Phi(A)}{\partial \text{vec} A^T},$$

i.e., the Jacobian matrix of $\text{vec}(\Phi)$. One useful result we shall use is the following [Neudecker, 1968]: If $d\Phi(A) = \sum_i M_i (dA) N_i$ for suitable matrices M_i, N_i, then the derivative is $D\Phi = \sum_i N_i^T \otimes M_i$. We define the commutation matrix $I_{(k,l)}$ as the square $kl \times kl$ block matrix partitioned into $k \times l$ blocks whose (i,j)'th block is 1 in the (j,i)'th coordinate and 0 everywhere else. Our goal is to use the delta method to determine the asymptotic distribution of the left side of $\hat{\Gamma}_m$.

Lemma 2 (Delta method)**.** *Let $(x_n)_{n \in \mathbb{N}} \subset \mathbb{R}^d$ be a sequence of random vectors such that $\sqrt{n}(x_n - x) \to_w \mathcal{N}(0, \Sigma)$ for some $x \in \mathbb{R}^d$ and a positive definite covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$. Assume furthermore that $h : \mathbb{R}^d \to \mathbb{R}^p$ is a continuous function that is continuously differentiable in a neighbourhood of x with Jacobian matrix $J = \frac{\partial h}{\partial y^T}|_{y=x}$. Then, $\sqrt{n}(h(x_n) - h(x)) \to_w \mathcal{N}(0, J\Sigma J^T)$.*

Furthermore, we need the following results on the sample covariance matrix.

Lemma 3. *Define $\tilde{X}_t = \begin{pmatrix} \Delta X_t^T & X_{1t-1}^T \end{pmatrix}^T$ and consider the sample covariance matrix $S_{\tilde{X}\tilde{X}}$. Then $\tilde{X}_t$ is stationary with,*

$$S_{\tilde{X}\tilde{X}} \to_p \Sigma_{\tilde{X}} = \begin{pmatrix} \Sigma_{\Delta X}^{11} & \Sigma_{\Delta X}^{12} & \beta^T \alpha \Sigma_X^{11} \\ \Sigma_{\Delta X}^{21} & \Sigma_{\Delta X}^{22} & 0 \\ \Sigma_X^{11} \alpha^T \beta & 0 & \Sigma_X^{11} \end{pmatrix}.$$

Let κ_{ijkl} be the joint cumulant of $U_{t,i}, U_{t,j}, U_{t,k}$, and $U_{t,l}$. If κ_{ijkl} vanishes for all $1 \leq i, j, k, l \leq p$, then $\sqrt{T}\text{vec}(S_{\tilde{X}\tilde{X}} - \Sigma_{\tilde{X}}) \to_w \mathcal{N}(0, \Xi)$, where

$$\Xi = \sum_{k=-\infty}^{\infty} \gamma_k \otimes \gamma_k + I_{(p+r,p+r)} \sum_{k=-\infty}^{\infty} \gamma_k \otimes \gamma_k \tag{2.A.5}$$

and $\gamma_k = \mathbb{E}(\tilde{X}_0 \tilde{X}_k^T)$.

Note that the second part of Lemma 3 holds also if we replace $\tilde{X}_t$ with any multivariate stationary linear process with vanishing fourth order cumulants. In fact, it is sufficient to assume that the cumulants are finite, but this complicates the expression for the asymptotic covariance somewhat so we keep the assumption. It holds specifically when U_t is Gaussian.

Proof of Lemma 3. $\tilde{X}_t$ is clearly stationary since ΔX_t and X_{1t-1} are stationary. From (2.2.3) and (2.2.4) we see that $\tilde{X}_t = \sum_{s=0}^{\infty} \Psi_s U_{t-s}$ where

$$\Psi_0 = \begin{pmatrix} I_r & 0_{r\times n} \\ 0_{n\times r} & I_n \\ 0_{r\times r} & 0_{r\times n} \end{pmatrix}, \qquad \Psi_s = \begin{pmatrix} \beta^T\alpha(I_r + \beta^T\alpha)^{s-1} & 0_{r\times n} \\ 0_{n\times r} & 0_{n\times n} \\ (I_r + \beta^T\alpha)^{s-1} & 0_{r\times n} \end{pmatrix} \text{ for } s \geq 1$$

and it is easily verified that $\sum_{s=0}^{\infty} ||\Psi_s|| < \infty$. The first part of the statement then follows from known results about linear processes (see e.g. Proposition C.12 in Lütkepohl [2005]).

Under the assumptions of Lemma 3, $\sqrt{T}(S_{\tilde{X}\tilde{X}} - \Sigma_{\tilde{X}})$ converges in distribution to some random matrix, N, with $\text{vec}(N)$ normal and covariance given by (see e.g. Roy [1989])

$$\text{Cov}(N_{i,j}, N_{k,l}) = \sum_{u=-\infty}^{\infty} (\gamma_u)_{ik}(\gamma_u)_{jl} + \sum_{u=-\infty}^{\infty} (\gamma_u)_{jk}(\gamma_u)_{il}.$$

Now let $\eta(i,j) = (p+r)(j-1) + i$ and observe that

$$\begin{aligned} (\gamma_u)_{ik}(\gamma_u)_{jl} &= (\gamma_u \otimes \gamma_u)_{\eta(i,j),\eta(k,l)}, \\ (\gamma_u)_{jk}(\gamma_u)_{il} &= \big((\gamma_u \otimes \gamma_u) I_{(p+r,p+r)}\big)_{\eta(i,j),\eta(k,l)}. \end{aligned}$$

Since $\Xi_{\eta(i,j),\eta(k,l)} = \text{Cov}(N_{i,j}, N_{k,l})$, this is exactly what we need to show, keeping in mind that $I_{(p+r,p+r)}(\gamma_u \otimes \gamma_u) = (\gamma_u \otimes \gamma_u) I_{(p+r,p+r)}$ [Magnus and Neudecker, 1979]. □

We now have all the tools we need to derive the asymptotics of $\hat{\Gamma}_m^{\cdot 1}$. For a matrix $M \in \mathbb{R}^{(p+r)\times(p+r)}$ write it in block form

$$M = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}$$

where M_{11} is $p \times p$ and M_{22} is $r \times r$. Denote by $\rho_1 \geq \cdots \geq \rho_r$ the generalized eigenvalues sorted in decreasing order for the generalized eigenvalue problem given by

$M_{21}(M_{11})^{-1}M_{12}$ and M_{22}. Let $v_1, ..., v_r$ be the corresponding generalized eigenvectors. Define the function $h : \mathbb{R}^{(p+r)\times(p+r)} \to \mathbb{R}^{r\times r}$ by $h(M) = M_{12}\sum_{k=1}^{m} v_k v_k^T$. We can write

$$dM_{11} = (I_p, 0_{p\times r})\, dM \begin{pmatrix} I_p \\ 0_{r\times p} \end{pmatrix}, \quad dM_{12} = (I_p, 0_{p\times r})\, dM \begin{pmatrix} 0_{p\times r} \\ I_r \end{pmatrix},$$

$$dM_{21} = (0_{r\times p}, I_r)\, dM \begin{pmatrix} I_p \\ 0_{r\times p} \end{pmatrix}, \quad dM_{22} = (0_{r\times p}, I_r)\, dM \begin{pmatrix} 0_{p\times r} \\ I_r \end{pmatrix}.$$

Also, we have $dM_{11}^{-1} = -M_{11}^{-1}(dM_{11})M_{11}^{-1}$ (see e.g. Thm. 8.3 in Magnus and Neudecker [2019]) whence

$$\begin{aligned} d(M_{21}M_{11}^{-1}M_{12}) = {} & (0_{r\times p}, I_r)\, dM \begin{pmatrix} M_{11}^{-1}M_{12} \\ 0_{r\times p} \end{pmatrix} \\ & - \left(M_{21}M_{11}^{-1}, 0_{p\times r}\right) dM \begin{pmatrix} M_{11}^{-1}M_{12} \\ 0_{r\times p} \end{pmatrix} + \left(M_{21}M_{11}^{-1}, 0_{p\times r}\right) dM \begin{pmatrix} 0_{p\times r} \\ I_r \end{pmatrix}. \end{aligned}$$

For ease of notation we now write $P_i = G_{11}e_i(G_{11}e_i)^T$ with e_i being the i'th unit vector, that is, $G_{11}e_i$ is the i'th column of G_{11}, $\Sigma_{X\Delta X} = (\Sigma_X^{11}\alpha^T\beta, 0_{r\times n})$, and $\Sigma_{\Delta XX} = \Sigma_{X\Delta X}^T$. Under Assumption 3, the map $M \mapsto v_i v_i^T$ is differentiable at $M = \Sigma_{\tilde{X}}$ (see Appendix 2.C). Let $\xi_i = D(v_i v_i^T)|_{M=\Sigma_{\tilde{X}}}$. Lemma 3 yields

$$\xi_i = \sum_{j\neq i}(\lambda_i - \lambda_j)^{-1}(P_i \otimes P_j + P_j \otimes P_i)F_i - (0_{r\times p}, P_i) \otimes (0_{r\times p}, P_i) \tag{2.A.6}$$

where

$$F_i = \left(\Sigma_{X\Delta X}\Sigma_{\Delta X}^{-1} \otimes \left(-\Sigma_{X\Delta X}\Sigma_{\Delta X}^{-1}, I_r\right), I_r \otimes \left(\Sigma_{X\Delta X}\Sigma_{\Delta X}^{-1}, -\lambda_i I_r\right)\right).$$

Then,

$$\xi = Dh|_{M=\Sigma_{\tilde{X}}} = \sum_{i=1}^{m} \begin{pmatrix} 0_{r\times p} & P_i \end{pmatrix} \otimes \begin{pmatrix} I_p & 0_{p\times r} \end{pmatrix} + (I_r \otimes \Sigma_{\Delta XX})\xi_i. \tag{2.A.7}$$

Observe that (2.A.1) also holds with $o_P(1)$ replaced by $o_P(T^{-\frac{1}{4}})$ and a similar argument as that applied in the proof of Theorem 2 therefore shows that $\hat{G}_{11}^{:m}(\hat{G}_{11}^{:m})^T = G_{11}^{:m}(G_{11}^{:m})^T + o_P(T^{-\frac{1}{2}})$. In particular,

$$\sqrt{T}\begin{pmatrix} \hat{\Gamma}_m^{11} - \Gamma_{11} - b \\ \hat{\Gamma}_m^{21} - \Gamma_{21} \end{pmatrix} = \sqrt{T}(h(S_{\tilde{X}\tilde{X}}) - h(\Sigma_{\tilde{X}})) + o_P(T^{-\frac{1}{2}})$$

and it is a straightforward application of Lemma 2 and Lemma 3 to prove that $\sqrt{T}\mathrm{vec}(h(S_{\tilde{X}\tilde{X}}) - h(\Sigma_{\tilde{X}})) \to_w \mathcal{N}(0, \xi\Xi\xi^T)$. Thus, we have identified the asymptotic distribution of the two left blocks. As we shall see below there is a much simpler expression for the asymptotic covariance matrix of $\sqrt{T}\mathrm{vec}(\hat{\Gamma}_m^{21} - \Gamma_{21})$. We are now ready to prove Theorem 3.

Proof of Theorem 3. Starting as in the proof of Theorem 1 and replacing $\hat{G}_{11}^{:m}(\hat{G}_{11}^{:m})^T$ with $G_{11}^{:m}(G_{11}^{:m})^T$ in the appropriate places, we find that

$$\begin{aligned}
\hat{\Gamma}_m^{11} &= S_{\Delta XX}^{11} G_{11}^{:m}(G_{11}^{:m})^T + o_P(T^{-\frac{1}{2}}) \\
\hat{\Gamma}_m^{21} &= S_{UX}^{21} G_{11}^{:m}(G_{11}^{:m})^T + o_P(T^{-\frac{1}{2}}) \\
\hat{\Gamma}_m^{12} &= T^{-1}\beta^T \alpha \Sigma_{XX} G_{11}^{:m}(G_{11}^{:m})^T H_1^T + o_P(T^{-1}) \\
\hat{\Gamma}_m^{22} &= o_P(T^{-1})
\end{aligned}$$

We derived the asymptotic distribution for the first two expressions above. The other two follow directly from Lemma 1.

The second result is also an easy consequence of Lemma 1, since $T^{\frac{1}{2}}\text{vec}(\hat{\Gamma}_m^{21} - \Gamma_{21})$ converges in distribution to $(G_{11}^{:m}(G_{11}^{:m})^T \otimes I_n)\text{vec}(V_{21})$, which, of course, is normal with mean 0 and covariance matrix as given in the theorem. Another way to arrive at the same result is to first observe that

$$\text{vec}\left(\hat{\Gamma}_m^{21} - \Gamma_{21}\right) = (I_r \otimes (0_{n\times r}, I_n))\,\text{vec}\begin{pmatrix}\hat{\Gamma}_m^{11} - \Gamma_{11} - b \\ \hat{\Gamma}_m^{21} - \Gamma_{21}\end{pmatrix}$$

and the asymptotic covariance of $\sqrt{T}\text{vec}(\hat{\Gamma}_m^{21} - \Gamma_{21})$ must therefore equal $\xi_{21}\Xi\xi_{21}^T$ where $\xi_{21} = (I_r \otimes (0_{n\times r}, I_n))\xi$. We compute

$$\xi_{21} = \left(0_{r\times p}, G_{11}^{:m}(G_{11}^{:m})^T\right) \otimes (0_{n\times r}, I_n, 0_{n\times r})$$

and thus $\xi_{21}(\gamma_k \otimes \gamma_k) = (\gamma_k \otimes \gamma_k)\xi_{21}^T = 0$ for all $k \neq 0$. Furthermore, we find that $\xi_{21} I_{(p+r,p+r)}(\gamma_0 \otimes \gamma_0)\xi_{21}^T = 0$ and

$$\xi_{21}(\gamma_0 \otimes \gamma_0)\xi_{21}^T = \xi_{21}(\Sigma_{\tilde{X}} \otimes \Sigma_{\tilde{X}})\xi_{21}^T = G_{11}^{:m}(G_{11}^{:m})^T \Sigma_X^{11} G_{11}^{:m}(G_{11}^{:m})^T \otimes \Sigma_U^{22}$$

which then results in the same covariance as before. □

When $m = r$ the expression for ξ simplifies significantly. Indeed, as noted after Lemma 3 in Appendix 2.C, we find that $\sum_{i=1}^r \xi_i = -(\Sigma_X^{11})^{-1} \otimes (\Sigma_X^{11})^{-1}$ and thus

$$\xi = \begin{pmatrix}0_{r\times p} & (\Sigma_X^{11})^{-1}\end{pmatrix} \otimes \begin{pmatrix}I_p & -\Sigma_{\Delta XX}(\Sigma_X^{11})^{-1}\end{pmatrix}.$$

We then compute $\xi(\gamma_k \otimes \gamma_k)\xi^T = \xi I_{(p+r,p+r)}(\gamma_k \otimes \gamma_k)\xi^T = 0$ for $k \neq 0$. For $k = 0$ we have $\gamma_0 = \Sigma_{\tilde{X}}$ and $\xi I_{(p+r,p+r)}(\gamma_0 \otimes \gamma_0)\xi^T = 0$. Thus,

$$\xi\Xi\xi^T = \xi(\Sigma_{\tilde{X}} \otimes \Sigma_{\tilde{X}})\xi^T = (\Sigma_X^{11})^{-1} \otimes \Sigma_U$$

which is the covariance matrix of $\text{vec}(V(\Sigma_X^{11})^{-1})$, i.e., our result is in line with the one derived in Theorem 1.

The following proof is an easy consequence of the discussion in Section 2.3.4.

Proof of Theorem 4. Assume that $k \geq r$. The first statement is then a direct consequence of (2.3.14), (2.3.15) and (2.3.17), and the fact that $T^{\frac{1}{2}}\text{vec}(\hat{\Pi}_k - \Pi) = (Q^T \otimes Q^{-1})T^{\frac{1}{2}}\text{vec}(\hat{\Gamma}_k - \Gamma)$, recalling that the two right blocks are $o_P(T^{-\frac{1}{2}})$.

For the second part, assume that $1 \leq k < r$. Then,

$$\hat{\Pi}_k - \Pi = Q^{-1}(\hat{\Gamma} - \Gamma)Q \to_p \alpha(\beta^T\alpha)^{-1}b\beta^T = \tilde{b}.$$

Applying the same argument as above and referring to the proof of Theorem 2, we obtain the desired distribution. □

Proof of Theorem 5. To prove consistency simply observe that

$$\hat{\Gamma}_{w_T} - \hat{\Gamma}_r = S_{\Delta XX}\left(\sum_{i=1}^{r}(w_{T,i} - 1)\hat{g}_i\hat{g}_i^T + \sum_{i=r+1}^{p} w_{T,i}\hat{g}_i\hat{g}_i^T\right).$$

$S_{\Delta XX}\hat{g}_i\hat{g}_i^T$ is bounded in probability for $i \leq r$ and converges in probability to 0 for $i > r$ from which it follows that both terms on the right hand side converge in probability to 0 for T going to infinity. The result then follows since $\hat{\Gamma}_r$ is consistent.

Now, for $0 \leq k \leq p$, define the matrices

$$D = \begin{pmatrix} T^{\frac{1}{2}}I_r & 0 \\ 0 & TI_n \end{pmatrix}, \quad B_k = \begin{pmatrix} b_k & 0 \\ 0 & 0 \end{pmatrix}, \quad B_w = \begin{pmatrix} b_w & 0 \\ 0 & 0 \end{pmatrix}$$

where b_k is the asymptotic bias of $\hat{\Gamma}_k$ for $0 \leq k \leq p-1$ and 0 otherwise. Let $W_{T,i} = w_{T,i+1} - w_{T,i}$ for $1 \leq i \leq p-1$, $W_{T,0} = w_{T,1}$, and $W_{T,p} = 1 - w_{T,p}$ and define W_i analogously for w instead of w_T so that, by assumption, $T(W_{T,i} - W_i)$ converges in probability to 0 for $T \to \infty$. Furthermore, define the random matrices $Z_0, ..., Z_p \in \mathbb{R}^{p\times p}$ such that $Z_0 = 0$, Z_k is the right-hand side of (2.3.17) for $1 \leq k < r$, Z_r is the right-hand side of (2.3.14), and Z_k is the right-hand side of (2.3.15) for $r < k \leq p$. It then follows from Theorems 1, 2, and 3 along with the continuous mapping theorem that

$$\left(\hat{\Gamma}_{w_T} - \Gamma - B_w\right)D = \sum_{k=0}^{p} W_{T,k}\left(\hat{\Gamma}_k - \Gamma - B_k\right)D \to_w \sum_{k=0}^{p} W_k Z_k$$

for $T \to \infty$. Upon rewriting the right-hand side of the above expression we obtain (2.4.21). □

2.B. Multiple Lags

In this section we consider processes of higher order. Let $d \geq 1$ and $\{Y_t\}_{t=0}^{\infty} \subset \mathbb{R}^p$ be an AR(d)-process. Similar to (2.1.1), the dynamics of Y_t can be expressed in VECM form by

$$\Delta Y_t = \Pi Y_{t-1} + \sum_{i=1}^{d-1}\Psi_i\Delta Y_{t-i} + Z_t$$

where $\{Z_t\}_{t=0}^{\infty}$ is a sequence of i.i.d. copies of Z_0 with 0 mean and finite fourth moment. Define the processes $X_{0t} = \Delta Y_t$, $X_{1t} = Y_{t-1}$, and $X_{2t} = (\Delta Y_{t-1}, ..., \Delta Y_{t-d})$ as well as the new parameter $\Psi = (\Psi_1, ..., \Psi_{d-1})$. We can then rewrite the equation as

$$X_{0t} = \Pi X_{1t} + \Psi X_{2t}$$

Assumptions similar to Assumptions 1 and 2 are needed to ensure a cointegrated process. We assume that $1 \leq r < p$.

Assumption 4. *The polynomial* $z \mapsto |(1-z)I_p - \Pi z - \sum_{i=1}^{d-1} \Psi_i (1-z) z^i|$ *has* $n = p - r$ *unit roots and all other roots are outside the unit circle.*

As before, this assumption implies that the rank of Π is $p - n = r$ so that we can write $\Pi = \alpha\beta^T$ for $\alpha, \beta \in \mathbb{R}^{p \times r}$ of full rank r.

Assumption 5. *The matrix* $\alpha_\perp^T (I_p - \sum_{i=1}^{d-1} \psi_i) \beta_\perp$ *is non-singular.*

The parameters are usually estimated as follows [Johansen, 1995]: First we find the residuals obtained from regressing X_{0t}, X_{1t}, and Z_t on X_{2t} denoted by R_{0t}, R_{1t}, and U_t, respectively. The reduced rank estimator, $\hat{\Pi}_k$ of Π, is then obtained as above starting with the equation

$$R_{0t} = \Pi R_{1t} + U_t.$$

After finding $\hat{\Pi}_k$ an estimator for Ψ is given by ordinary least squares, i.e., $\hat{\Psi}_{LS}$ is obtained by regressing X_{2t} on $X_{0t} - \hat{\Pi}_k X_{1t}$. The asymptotics in this case can be derived from the previous section. Indeed, as shown in Johansen [1995], similar limiting results as those given in Lemma 1 exist for the empirical cross-covariances given by R_{0t} and R_{1t} and the limiting behaviour of $\hat{\Psi}_{LS}$ is studied in the usual way.

2.C. Auxiliary results

We state here some results from perturbation theory of linear operators. These will be relevant especially for proving convergence of eigenvectors and eigenvalues. For more information, see Kato [2013]. Let $M, N \in \mathbb{R}^{p \times p}$ and denote by $||\cdot||_F$ the Frobenius-norm. Define $\rho(M, N) \in \mathbb{C}^p$ to be the ordered p-tuple that contains the solutions to $M - \rho N = 0$ counted with multiplicity.

Lemma 1. *Assume that* N *is non-singular. Then the map* $(M, N) \mapsto \rho(M, N)$ *is continuous in the sense that for a sequence* $M_n, N_n \in \mathbb{R}^{p \times p}$ *with* $||M - M_n||_F + ||N - N_n||_F \to 0$ *it holds that*

$$||\rho(M, N) - \rho(M_n, N_N)|| \to 0.$$

Proof. This is Theorem 5.14 in Kato [2013] after observing that for $\rho \in \rho(M, N)$

$$|N^{-1}M - I\rho| = 0$$

□

If M and N are real symmetric, then $\rho(M,N) \in \mathbb{R}^p$. Writing $\rho(M,N)_i = \rho_i$ there then exist real-valued vectors $v_1, ..., v_p$ satisfying $Mv_i = \rho_i N v_i$ and $v_i^T N v_i = \delta_{ij}$ for $i, j = 1, ..., p$. We call v_i the generalized eigenvector corresponding to ρ_i. The generalized eigenvectors are not unique, but we can define $P_i(M,N) := \sum_{j:\rho_i=\rho_j} v_i v_i^T$ for $i = 1, ..., p$ which is then uniquely determined by M, N. Note that for $\rho_i = \rho_j$ we will also have $P_i = P_j$. If we denote by S the space of real symmetric $p \times p$ matrices, we are led to define the maps from $S \times S$ to $\mathbb{R}^{p\times p}$ given by P_i for $i = 1, ..., p$. The following two lemmas concern the smoothness of these maps.

Lemma 2. *Let S_+ denote the space of real positive definite $p \times p$ matrices. P_i is continuous at points $(M,N) \in S \times S_+$. This means that if $(M_n, N_n) \in \mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p}$ with $||M_n - M||_F + ||N_n - N||_F \to 0$ for $n \to \infty$, then*

$$||P_i(M_n, N_n) - P_i(M,N)||_F \to 0.$$

Proof. We may assume for n large enough that N_n is positive definite since it converges towards a positive definite matrix. Define $S = N^{-\frac{1}{2}} M N^{-\frac{1}{2}}$ and $S_n = N_n^{-\frac{1}{2}} M_n N_n^{-\frac{1}{2}}$. Clearly the eigenvalues of S are $\rho_1, ..., \rho_p$ with the corresponding orthonormal eigenvectors given by $\tilde{v}_i = N^{\frac{1}{2}} v_i$ for $i = 1, ..., p$ and similarly for S_n. The result then follows from Theorem 2.23 and 3.16 in Kato [2013]. □

Lemma 3. *Assume $(M,N) \in S \times S_+$ with simple eigenvalues, i.e., $\rho_1 > ... > \rho_p$. Then P_i and ρ_i are continously differentiable at (M,N). Furthermore, the differential of P_i is given by*

$$dP_i = -P_i(dN)P_i - (M - \rho_i N)^+ (dM - \rho_i dN)P_i - P_i(dM - \rho_i dN)(M - \rho_i)^+$$

where $(M - \rho_i N)^+ = \sum_{j:\rho_j \neq \rho_i} (\rho_j - \rho_i)^{-1} P_j$.

Proof. The first statement follows directly by Theorem 8.9 in Magnus and Neudecker [2019] after transforming the problem as above. To find the expression for the differential, we start with the defining equations

$$MP_i = \rho_i N P_i, \qquad P_i N P_i = P_i.$$

The first equation yields $(M - \rho_i N)dP_i = -(dM - \rho_i dN)P_i + (d\rho_i)NP_i$. Note that $P_j N P_i = \delta_{ij} P_i$ and that $M - \rho_i N = (\sum_{j=1}^p (\rho_j - \rho_i) N P_j) N$. Multiplying the above differential equation by $(M - \rho_i N)^+$ on each side therefore yields

$$(I_p - P_i N)dP_i = -(M - \rho_i N)^+ (dM - \rho_i dN)P_i, \tag{2.C.1}$$
$$(dP_i)(I_p - NP_i) = -P_i(dM - \rho_i dN)(M - \rho_i N)^+. \tag{2.C.2}$$

Now after introducing differentials into the equation $P_i N P_i = P_i$ we arrive at $(I_p - P_i N)dP_i + (dP_i)(I_p - NP_i) = dP_i + P_i(dN)P_i$. Plugging this into (2.C.1) + (2.C.2) gives us exactly the equation stated in the Lemma. □

It is not too hard to show that

$$\sum_{i=1}^{p}(M-\rho_i N)^+(dM-\rho_i dN)P_i = -\sum_{i=1}^{p}P_i(dM-\rho_i dN)(M-\rho_i N)^+$$

and from the previous Lemma we then obtain for $P=\sum_{i=1}^{p}P_i$ that

$$dP = -\sum_{i=1}^{p}P_i(dN)P_i.$$

But we also have that $P=N^{-1}$ from which it follows that $dP=-P(dN)P$ and thus $-\sum_{i=1}^{p}P_i(dN)P_i=-P(dN)P$ and so the differential of P simplifies significantly.

2.D. Simulation study

Here we describe in detail the simulation experiments from Section 2.5. The simulations were run in Python 3.9.1 and the code can be found on GitHub.[2]

2.D.1 Comparison of Estimators

We consider an array of parameters $\Gamma_c \in \mathbb{R}^{p\times p}$ for $p=3,6,9$ and $c\in\{0,0.2,0.4,...,29.8,30\}$. For each c and p, we let Γ_c be the diagonal matrix given by

$$(\Gamma_c)_{ii} = \begin{cases} -1.5 & \text{if } i \le p/3, \\ -c/T & \text{if } p/3 < i \le 2p/3 \\ 0 & \text{otherwise.} \end{cases}$$

Throughout we fix $T=100$. For each Γ_c we draw $X_1,...,X_T$ and X_{T+1} and compute the MSPE across 4 million simulations. The results are given in Figure 2.D.1.

To gain some insight we also plotted the mean and standard deviation of the individual weights across all simulations for $p=3$ (see Figures 2.D.2 and 2.D.3). Clearly, choosing smoother weights significantly reduces the variance of the weights for eigenvectors where the corresponding eigenvalue is small. This is particularly apparent for $\hat{w}_2(0.1,0.1)$. While on average the weight is similar to $\hat{w}$, it is not as steep for increasing c and its variance is significantly lower.

Finally, we decomposed the lines in Figure 2.5.2 into bias and variance, the results are plotted in Figure 2.D.4 and 2.D.5. For an estimator $\hat{\Gamma}$ of the true parameter Γ, bias here refers to the real number $\|\mathbb{E}\hat{\Gamma}-\Gamma\|^2$ and the variance refers to $\mathrm{tr}(\mathrm{Var}(\Gamma))$.

2.D.2 Comparison of Distributions

We generated 1000 i.i.d. samples of length $T=5000$ of the process $Y_t\in\mathbb{R}^4$ given by (2.1.1). For each sample, the errors are i.i.d. $Z_t\sim\mathcal{N}(0,\Sigma_Z)$ where Σ_Z is a random

[2] https://github.com/cholberg/coint_CLT

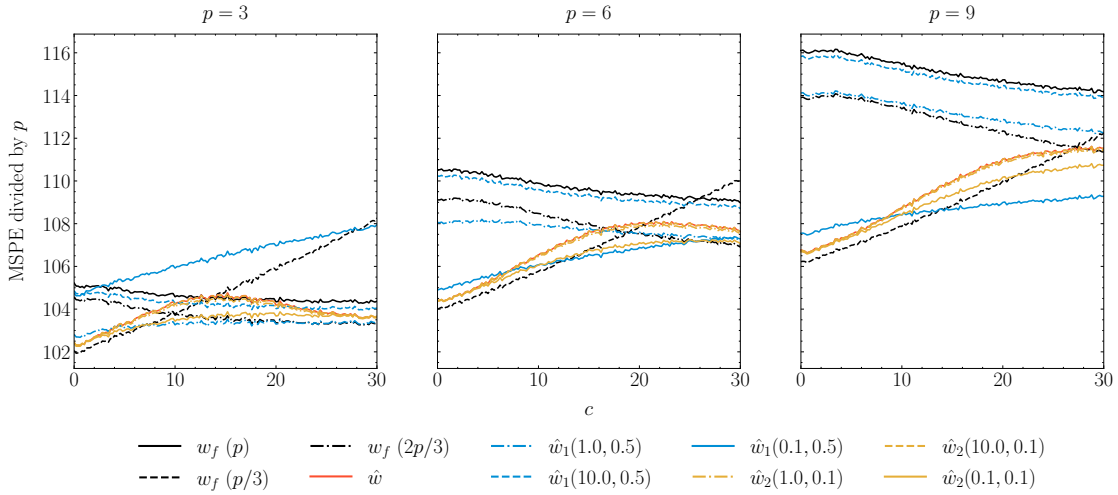

Figure 2.D.1: Mean square prediction error (MSPE) of different weighted reduced rank estimators for varying dimensions and $c \in [0, 30]$ where the underlying autoregressive matrix, Γ_c, has a third of its eigenvalues set to $-c/T$, a third set to 0 and a third set to $-3/2$. Sample size is fixed at $T = 100$.

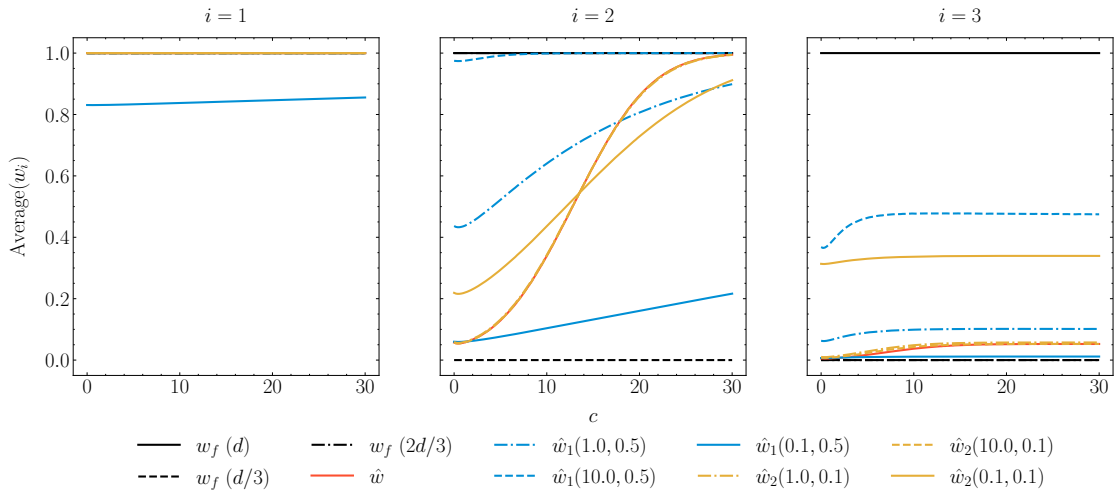

Figure 2.D.2: Mean of weights across 4 million simulations for $d = 3$ and $c \in [0, 1]$ where the underlying autoregressive matrix, Γ_c, has a third of its eigenvalues set to $-c/T$, a third set to 0 and a third set to $-3/2$. Sample size is fixed at $T = 100$.

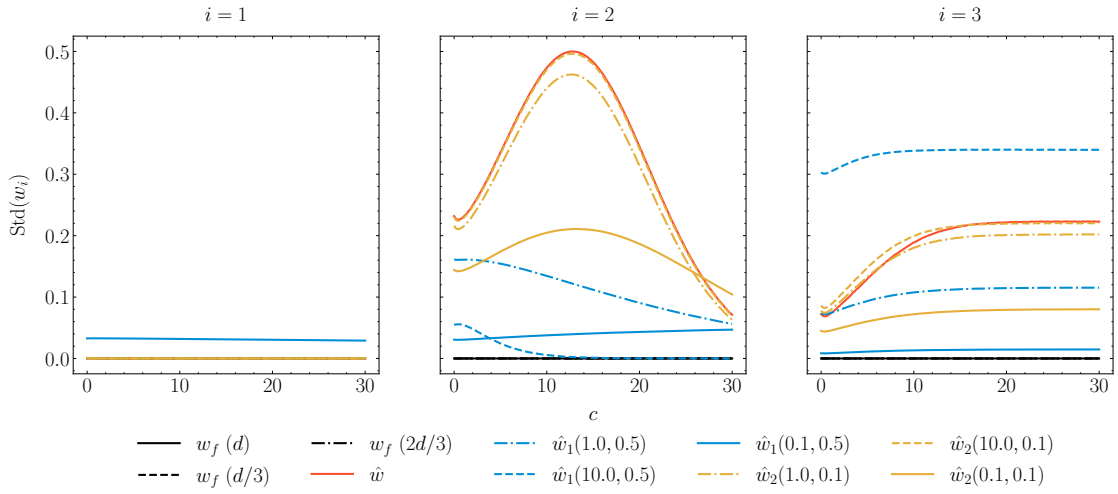

Figure 2.D.3: Standard deviation of weights across 4 million simulations for $d = 3$ and $c \in [0, 1]$ where the underlying autoregressive matrix, Γ_c, has a third of its eigenvalues set to $-c/T$, a third set to 0 and a third set to $-3/2$. Sample size is fixed at $T = 100$.

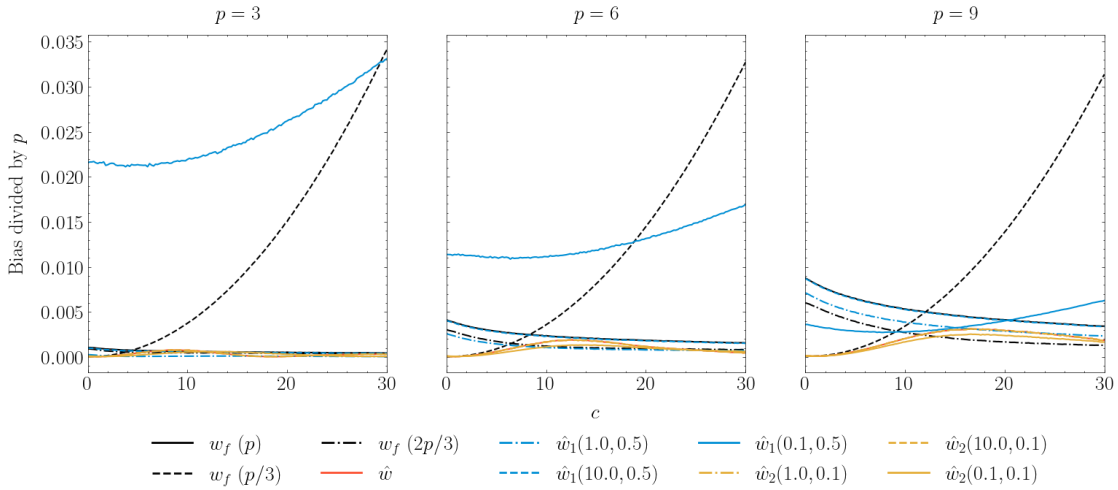

Figure 2.D.4: Bias of different weighted reduced rank estimators for varying dimensions and $c \in [0, 30]$ where the underlying autoregressive matrix, Γ_c, has a third of its eigenvalues set to $-c/T$, a third set to 0 and a third set to $-3/2$. Sample size is fixed at $T = 100$.

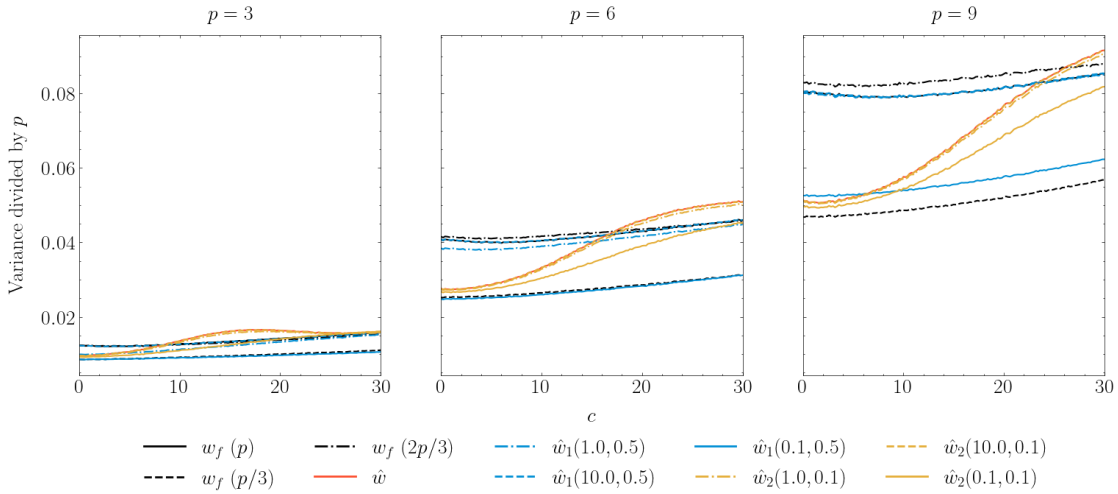

Figure 2.D.5: Variance of different weighted reduced rank estimators for varying dimensions and $c \in [0, 30]$ where the underlying autoregressive matrix, Γ_c, has a third of its eigenvalues set to $-c/T$, a third set to 0 and a third set to $-3/2$. Sample size is fixed at $T = 100$.

positive definite matrix that is given by

$$I_4 + \frac{1}{2} U U^T$$

with U_{ij} i.i.d. uniformly over $[0, 1]$ for $i, j = 1, ..., 4$. The matrix $\Pi \in \mathbb{R}^{4\times 4}$ is of rank 2 and can be decomposed as $\Pi = \alpha\beta^T$ where

$$\alpha = \begin{pmatrix} -0.7 & 0 \\ 0 & -0.7 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad \beta = \begin{pmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \\ 0 & 0 \end{pmatrix}.$$

This ensures that Assumptions 1, 2, and 3 are fulfilled with probability 1 so that the process is $I(1)$ and cointegrated. The cointegration rank is equal to the rank of Π and the cointegration relations are given by the columns of β. Each sample is Q-transformed to get X_t as in (2.2.2) and we computed the estimators $\hat{\Gamma}_1$, $\hat{\Gamma}_2$ and $\hat{\Gamma}_4$ to obtain the empirical large-sample distribution in the three different cases.

The asymptotic densities were obtained by generating 1000 samples from (2.3.14), (2.3.15), and (2.3.17) with the parameters as given above.

2.D.3 Rank selection vs. Bias

We now turn to the relation between the r non-zero eigenvalues in (2.A.4) and rank-selection. We consider a high-dimensional process $Y_t \in \mathbb{R}^{40}$ generated by (2.1.1) under different parameter settings. The parameters are chosen in such a way that $||\Pi||_F$

remains fixed in all settings with cointegration rank $r = 20$. However, the sequence of eigenvalues changes. For each setting, we consider different values of λ_{min} and λ_{max} such that $\lambda_{min} < \lambda_{max}$ correspond to the smallest and largest squared eigenvalue, respectively. In order to keep $||\Pi||_F$ fixed, an increase in λ_{min} leads to a decrease in λ_{max} so that it suffices to only specify the value of λ_{min}. Smaller values of λ_{min} represent cases in which there are many small eigenvalues so that we would expect the cointegration test to be more prone to underestimate the true rank. To be precise for a given choice of $\lambda_{min} \in \{0.01, 0.03, 0.1, 0.3\}$ the samples are generated as follows: For each sample the errors are i.i.d. $Z_t \sim \mathcal{N}(0, I_{40})$. We fix

$$\beta = \begin{pmatrix} I_{20} \\ 0_{20\times 20} \end{pmatrix}$$

and let $\lambda_{max} = 0.81 - \lambda_{min}$. This choice may seem arbitrary, but it basically ensures that the process is non-explosive and that the norm of Π does not depend on λ_{min} as argued below. Now define

$$\lambda_k = \lambda_{min} + \frac{(\lambda_{max} - \lambda_{min})(k-1)}{19}, \quad \text{for } k = 1, ..., 20.$$

Setting $D = \text{diag}(\sqrt{\lambda_1}, ..., \sqrt{\lambda_{20}})$ and letting

$$\alpha = \begin{pmatrix} -2D \\ 0_{20\times 20} \end{pmatrix}$$

then ensures that Assumptions 1, 2, and 3 are fulfilled with cointegration rank 20 and the non-zero eigenvalues defined in (2.A.4) being exactly $\sqrt{\lambda_1}, ..., \sqrt{\lambda_{20}}$. Here, of course, $\Pi = \alpha\beta^T$ and $\Sigma_Z = I_{40}$. Observe that the process Y_t is already split up into its stationary and random walk part. By construction we also have that

$$||\Pi||_F^2 = \sum_{i=k}^{20} \lambda_k = 20\lambda_{min} + \frac{\lambda_{max} - \lambda_{min}}{19} \sum_{k=1}^{20}(k-1) = 10(\lambda_{min} + \lambda_{max}) = 8.1$$

so that $||\Pi||_F$ does not depend on our choice of λ_{min}.

Since the asymptotic distributions of the test statistics described in Section 2.4.1 are non-standard and non-applicable in dimensions that much exceed $p = 12$ we will instead use a bootstrap approach as described in Cavaliere et al. [2015] to estimate the rank. In all simulations we used $B = 299$ bootstrap samples for each test.

In Fig. 2.D.6 we estimate the probability of choosing a given rank for different values of λ_{min}. For each choice, we simulate 200 samples of Y_t of length $T = 200$ and estimate the rank using the sequential testing approach described above. We used the trace statistic and bootstrap to determine the asymptotic distribution under each hypothesis. In line with our expectations, the rank tends to be underestimated for smaller values of λ_{min}. It appears that an increase in λ_{min} causes a shift towards the true rank in the distribution of estimated ranks.

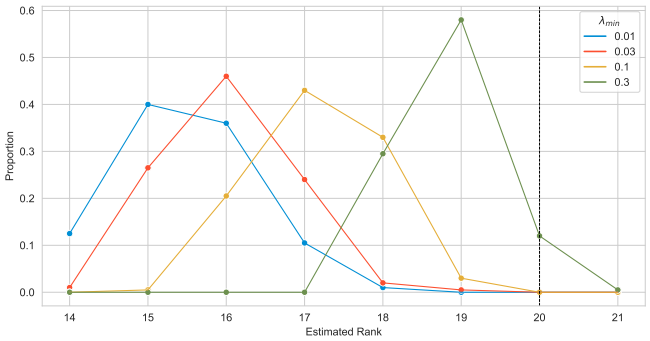

Figure 2.D.6: Distribution of the estimated rank under different parameter settings. The results are based on 200 simulations of $(Y_0, ..., Y_{200})$. The vertical line at 20 illustrates the true rank of Π. The dimension is $p = 40$.

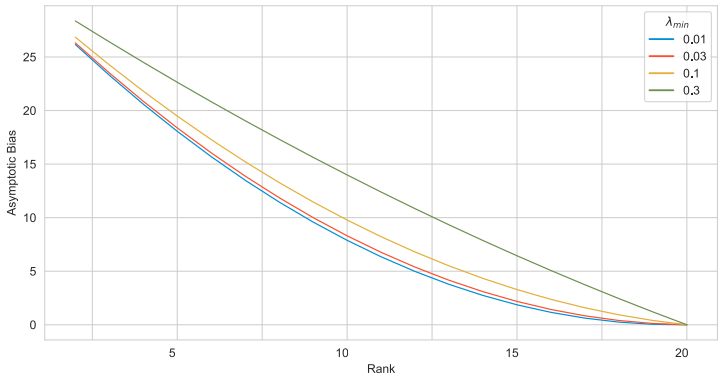

Figure 2.D.7: Asymptotic bias of $\hat{\Pi}_k$ for different values of $k \in \{2, ..., 20\}$ and choices of $\lambda_{min} \in \{0.01, 0.03, 0.1, 0.3\}$. We plot here the Frobenius norm of the bias b defined in (2.3.16).

When studying the asymptotic bias in the different cases, we see an adverse effect. For smaller values of λ_{min}, the bias tends to be very small as long as we only underestimate the true rank by a little. This is illustrated in Fig. 2.D.7. Choosing $k = 14$ in the case where $\lambda_{min} = 0.01$ leads to approximately the same asymptotic bias as choosing $k = 18$ in the case where $\lambda_{min} = 0.3$ even though in the former case we are quite far off from the true rank.

Uniform Inference for Cointegrated Vector Autoregressive Processes

Christian Holberg, Susanne Ditlevsen

Abstract

Uniformly valid inference for cointegrated vector autoregressive processes has so far proven difficult due to certain discontinuities arising in the asymptotic distribution of the least squares estimator. We extend asymptotic results from the univariate case to multiple dimensions and show how inference can be based on these results. Furthermore, we show that lag augmentation and a recent instrumental variable procedure can also yield uniformly valid tests and confidence regions. We verify the theoretical findings and investigate finite sample properties in simulation experiments for two specific examples.

2.8 Introduction

Persistence, i.e., long term sensitivity to small shocks, appears to be a commonly occurring characteristic of many stochastic systems encountered in practice. Such processes are often modeled with cointegration where the persistence can be attributed to a number of shared stochastic trends (random walks). Due to their relative simplicity, cointegration models are widely applied. Cointegration in vector-valued autoregressive processes arises when the characteristic polynomial possesses at least one unit root [Johansen, 1988, 1991, 1995]. The asymptotic theory and, hence, inference is heavily reliant upon the fact that a fixed number of roots can be assumed exactly one and the rest stay sufficiently far away from one. This is the case even for simple regression methods such as ordinary least squares. In practice, such assumptions are overly restrictive and more flexible models are often needed for a better description of the empirical data. Slight deviations from the unit root assumption can severely deteriorate the results of the statistical analysis [Elliott, 1998]. Thus, the need arises for inference methods that are uniformly valid over a range of stationary and non-stationary behaviors.

So far, uniform guarantees have only been provided for methods of inference in the univariate case. Bootstrap inference algorithms are presented in Andrews [1993], Hansen [1999] with uniform guarantees given in Mikusheva [2007]. Furthermore, Mikusheva

[2007] provides a uniform asymptotic framework for one-dimensional autoregressive processes with potential unit roots, which served as an inspiration for much of the work in this paper.

The problem is well-understood in one dimension, but less progress has been made in multiple dimensions. This deficiency is certainly not due to a lack of methods being proposed. Indeed, multiple methods which are supposedly robust against local deviations to the unit root assumption exist in the literature. One prominent strand employs lag augmentation [Dolado and Lütkepohl, 1996, Toda and Yamamoto, 1995]. The idea is simple and easy to apply, but lacks efficiency since it essentially involves overfitting the model. While these results are in a sense stated as uniform, this is never made mathematically precise. In particular, it is not stated over which parameter space, the procedures will yield tests of uniform size (or coverage probability in the case of confidence intervals). Other methods, seeking to avoid the inefficiency problem, derive their methods under specific configurations of parameters. The main approach is to assume that the roots are all of a similar proximity to one. In particular, the autoregressive matrix is modeled as a sequence of matrices that approach the identity matrix at some given rate k_n. This is, for example, the setup in the instrumental variable methodology (IVX) developed in Kostakis et al. [2015], Magdalinos and Phillips [2020], Phillips et al. [2009]. In these cases, however, asymptotic guarantees are only provided for the given sequence of parameters. Furthermore, a lot of focus has been given to predictive regression problems in which the predictive power of the past of one process on the future of another is assessed. Efficient tests are developed in Campbell and Yogo [2006], Jansson and Moreira [2006] and, based on the ideas of uniform inference for univariate autoregressive processes, Phillips [2014] leverage these methods for uniformly valid inference. Unfortunately, most of the work only covers the bivariate case in which the regressor is a univariate autoregressive process so that the theory for one dimension directly applies. Another branch of research concerns inference on cointegrating relations robust to deviations from the unit root assumption [Duffy and Simons, 2023, Franchi and Johansen, 2017], but, again, with asymptotic results only covering specific sequences of parameters.[3]

Quite a bit of effort has gone into establishing asymptotic statements of the flavor that we seek here. Generic results on uniform convergence are provided in Andrews et al. [2020]. However, they hold only insofar as one can establish the right asymptotic distributions under arbitrary sequences of parameters.[4] While progress in this direction has been made [Phillips and Lee, 2013, 2015], there has thus far been no general answer to this problem. For example, these papers do not allow the process to have parts that

[3] In Duffy and Simons [2023], the general theory is developed without any restrictions other than a fixed number of roots being bounded away from unity. However, to study the large sample behavior of their proposed methods (Section 3.1.4), the authors impose the familiar local-to-unity framework where the autoregressive matrix is modeled as a sequence of matrices drifting towards the identity at rate n.

[4] Note that the definition of uniform convergence given in the present paper (see Def. 1) is equivalent to convergence along all sequences of parameters. Thus, establishing the validity of the necessary assumptions in Andrews et al. [2020] (in particular Assumption A1 and S) is essentially the goal of the work presented here.

remain stationary for all sample sizes, and only diagonal regression matrices are allowed. The most general result in this direction is given in Phillips [1988] where asymptotic approximations are established that hold generically for drifting autoregressive matrices of the form $\Gamma = \exp(Cn^{-1})$ with no further structure assumed on the matrix C. This setting, also called the local-to-unity setting, has received a lot of attention due to the challenges arising on the inferential side. As might have been expected, similar to the univariate case, considering the local-to-unity framework provides the link between the limiting behavior of unit root processes and purely stationary processes.

Extending the theory in Mikusheva [2007] to multiple dimensions, as is our goal, runs into several difficulties. Firstly, it is not clear what assumptions to put on the autoregressive matrix to ensure that the asymptotic results hold uniformly while still covering all relevant cases. In one dimension, the autoregressive parameter is a scalar and it is sufficient that it is real, bounded in norm by 1, and bounded away from -1 by some small δ. We need to extend this idea to matrices. A simple approach is to assume that at most one root is allowed to be close to unity as done in Mikusheva [2012]. However, in some applications this is too restrictive and it would be useful to allow for arbitrarily many roots close to one. Secondly, while many of the results on the asymptotics of the sample covariance matrices generalize nicely to multiple dimensions, the proofs are more involved. For example, Mikusheva [2007] uses Skorohod's embedding, which famously only works in one dimension, to prove that the errors can be assumed to be Gaussian. The third and perhaps most profound difficulty is that the multivariate setting allows for cointegrated systems (or almost cointegrated systems in the case where the roots are only close to unity). This gives rise to certain asymptotic discontinuities. In particular, it necessitates a proper normalization of the sample covariances. These problems extend to the inferential side where, additionally, computational challenges arise. Naively adapting, for example, the grid bootstrap approach of Hansen [1999] would cause the computational complexity to explode. On the other hand, in many settings one might have reason to believe that there is an upper bound to how many roots can fall close to unity. Restricting, as in Mikusheva [2012], all but one root away from the unit circle can simplify inference significantly (see also the remarks after Theorem 2).

What we seek here, then, is to precisely define the parameter space of interest, Θ,[5] and establish the validity of inferential procedures not just for certain specific regions or sequences in Θ, but uniformly over all of Θ. If we insist that this definition of the size of a test, i.e., the highest probability of rejecting the null across the entire parameter space, is the appropriate one, then the previous work has been insufficient in providing such assurance. The missing piece, no doubt, is providing uniform approximations of the two key covariance matrices given in (2.10.1), which is the main objective of the present work. With this result in hand, our hope is that the validity of most of the proposed methods can be established fairly easily. Indeed, a few initial steps in this direction are made in the latter half of the paper.

The main contributions are the following. First, we show that the results in Phillips

[5] This space being one that includes exactly all the VAR processes that might be integrated of order 1 and cointegrated.

[1988] can be proven to hold in the sense of uniform convergence (Definition 1) over an appropriate parameter space, which generalizes the weak limit of a local-to-unity sequence of parameters. Essentially, this means that the asymptotic distributions obtained under the local-to-unity assumption can be used to uniformly approximate statistics arising in the VAR model, under any arbitrary drifting sequences of parameters and even in cases where the process is stationary. In this sense, we do for the multivariate AR model what was done in Mikusheva [2007] for the scalar model.[6] The main result is Theorem 1, stating that the asymptotic distributions of the relevant sample covariances can be approximated by stochastic integrals of Ornstein-Uhlenbeck processes (a direct analog to the univariate case). Both the result and the proof are interesting in their own rights. As part of the proof we show that one can approximate the finite sample distribution of crucial statistics by replacing the general error terms with Gaussian errors. We can sample from this approximation at a comparably low computational cost which facilitates inference greatly.

Second, we give a few preliminary examples of how these results can be utilized to prove uniform validity of simple confidence regions for the autoregressive matrix based on the t-statistic. We also show that confidence regions constructed with IVX and lag augmentation are uniformly valid.

Third, we show how these confidence regions can be used to answer more general inference questions. The two main applications are confidence intervals for a single coordinate and predictive regression testing with a multivariate regressor. To the best of our knowledge, there have thus far been no attempts in the literature to deal with these applications in a uniform fashion (apart from lag augmentation). We run Monte Carlo experiments to verify the theoretical results and compare the finite sample properties of the different methods.

Our last contribution is the development of efficient algorithms to solve these inferential tasks. We show how the *Evaluation-Approximation-Maximization* (EAM) algorithm from Kaido et al. [2019] circumvents the exploding computational cost inherent in algorithms relying on grid-like methods. Combined with the Gaussian approximation results, this is what makes inference possible in our two main applications.

The paper is structured as follows. Section 2.9 introduces notation and relevant concepts. In particular, it explains the concept of uniform convergence of random variables and presents vector autoregressive processes. Section 2.10 is devoted to presenting and proving the main asymptotic results under the VAR(1) model. We extend to the general VAR(p) model in Section 2.11. Section 2.12 deals with inference and shows how the results of Section 2.10 can be applied to obtain uniformly valid confidence regions. Furthermore, it contains a section on predictive regression, lag augmentation, and IVX. Section 2.13 contains the results of our Monte Carlo experiments. Finally, Section 2.14 concludes. The Appendix contains proofs and further technical details on martingale limit results, the Gaussian approximation, the simulation experiments as well as the EAM algorithm, and details on lag augmentation and IVX.

[6] We also show that uniformity holds over a *family* of martingale difference error processes which is an additional generalization of Mikusheva [2007].

2.9 Preliminaries

Notation: For a matrix $A \in \mathbb{C}^{d\times d}$, A^T denotes its conjugate transpose and its trace is $\mathrm{tr}(A)$. We write $\sigma_{max}(A)$ ($\sigma_{min}(A)$) for the largest (smallest) singular value of A, and $\lambda_{max}(A)$ ($\lambda_{min}(A)$) for the eigenvalue of A with the largest (smallest) magnitude. $||A||$ is the Frobenius norm and $||A||_2$ is the spectral norm, i.e., $||A|| = \sqrt{\mathrm{tr}(A^T A)}$ and $||A||_2 = \sqrt{\sigma_{max}(A)}$. For vectors, $||\cdot||$ is the usual Euclidean norm. Define S_d to be the set of $d \times d$ positive semidefinite matrices. We employ the usual big-O and little-o notation and use o_p to denote convergence in probability and O_p to denote boundedness in probability.

2.9.1 Uniform convergence of random variables

The definitions of uniform convergence in probability and in distribution are essentially the same as in Kasy [2019], Lundborg et al. [2022]. Assume some background probability space, $(\Omega, \mathcal{F}, \mathbb{P})$, on which all future random variables are defined. For two random vectors, X and Y, taking values in $(\mathbb{C}^d, \mathcal{B}(\mathbb{C}^d))$, we denote by P_X and P_Y the law of X and Y and write $X \overset{\mathcal{L}}{=} Y$ if they are equal in law. Let BL_1 be the space of functions $f : \mathbb{C}^d \to [-1, 1]$ that are Lipschitz continuous with constant at most 1. Let $\mathcal{P}(\mathbb{C}^d, \mathcal{B}(\mathbb{C}^d))$ be the set of probability measures on $(\mathbb{C}^d, \mathcal{B}(\mathbb{C}^d))$. The bounded Lipschitz metric on $\mathcal{P}(\mathbb{C}^d, \mathcal{B}(\mathbb{C}^d))$ is given by

$$d_{BL}(\mu, \nu) := \sup_{f\in BL_1} \left| \int_{\mathbb{C}^d} f d\mu - \int_{\mathbb{C}^d} f d\nu \right|, \quad \mu, \nu \in \mathcal{P}(\mathbb{C}^d, \mathcal{B}(\mathbb{C}^d)).$$

We use the shorthand $d_{BL}(X, Y) = d_{BL}(P_X, P_Y)$ to denote the bounded Lipschitz metric between the laws of two random variables, X and Y. It is well known that d_{BL} metrizes weak convergence which motivates the following definition of uniform convergence.

Definition 1 (Uniform convergence)**.** Let $(X_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$ and $(Y_{n,\theta})_{n\in\mathbb{N},\theta\in\Theta}$ be two sequences of families of random d-dimensional vectors defined on $(\Omega, \mathcal{F}, \mathbb{P})$ and indexed by some set Θ (of possibly infinite dimension).

1. We say that $X_{n,\theta}$ *converges uniformly to* $Y_{n,\theta}$ *over* Θ *in distribution* (or, for short, $X_{n,\theta} \to_w Y_{n,\theta}$ uniformly over Θ) if
$$\lim_{n\to\infty} \sup_{\theta\in\Theta} d_{BL}\left(X_{n,\theta}, Y_{n,\theta}\right) = 0.$$
2. We say that $X_{n,\theta}$ *converges uniformly to* $Y_{n,\theta}$ *over* Θ *in probability* (or, for short, $X_{n,\theta} \to_p Y_{n,\theta}$ uniformly over Θ) if, for every $\epsilon > 0$,
$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \mathbb{P}\left(||X_{n,\theta} - Y_{n,\theta}|| > \epsilon\right) = 0.$$

Uniform convergence could also be stated as convergence along all sub-sequences $\theta_n \subset \Theta$ (see Definition 2 and Lemma 1 in Kasy [2019]). Additionally, we allow the limiting distribution to be a sequence, since the results below are stated in terms of an approximating sequence of random variables. We obtain the conventional notion by letting $Y_{n,\theta} = Y_\theta$.

2.9.2 Model

Consider some $\Theta \subset \mathbb{R}^{d\times d} \times S_d \times \mathbb{R}_+$. For any $\theta \in \Theta$ there exist $\Gamma_\theta, \Sigma_\theta$, and c_θ such that $\theta = (\Gamma_\theta, \Sigma_\theta, c_\theta)$. Let $N_\theta \in \{1, ..., d\}$ denote the number of distinct eigenvalues of Γ_θ and $\lambda_\theta \in \mathbb{C}^{N_\theta}$ the corresponding vector of ordered eigenvalues, that is, $|\lambda_{\theta,1}| \geq |\lambda_{\theta,2}| \geq ... \geq |\lambda_{\theta,N_\theta}|$ with multiplicities $m_{\theta,1}, ..., m_{\theta,N} \in \{1, ..., d\}$. Where this does not cause confusion, we omit the subscript θ. Let $(X_{t,\theta})_{t\in\mathbb{N},\theta\in\Theta}$ and $(\epsilon_{t,\theta})_{t\in\mathbb{N},\theta\in\Theta}$ be two families of $\mathbb{R}^d$-valued stochastic processes and $(\mathcal{F}_{t,\theta})_{t\in\mathbb{N}}$ the filtration generated by $(\epsilon_{t,\theta})_{t\in\mathbb{N}}$.

Assumption M. *$X_{t,\theta}$ and $\epsilon_{t,\theta}$ satisfy the following:*

M.1. $\epsilon_{t,\theta}$ is a stationary martingale difference sequence wrt. $\mathcal{F}_{t,\theta}$, that is,

$$\sup_{t\in\mathbb{N},\theta\in\Theta} \mathbb{E}||\epsilon_{t,\theta}|| < \infty$$

and $\mathbb{E}(\epsilon_{t,\theta}|\mathcal{F}_{t-1,\theta}) = \mathbb{E}\epsilon_{0,\theta} = 0$ for all $t \geq 1, \theta \in \Theta$.

M.2. For all $\theta \in \Theta$, the conditional covariance matrix of $\epsilon_{t,\theta}$ exists and is given by $\mathbb{E}(\epsilon_{t,\theta}\epsilon_{t,\theta}^T|\mathcal{F}_{t-1,\theta}) = \mathbb{E}\epsilon_{0,\theta}\epsilon_{0,\theta}^T = \Sigma_\theta$ a.s. for all $t \geq 1$.

M.3. There exists some small $\delta > 0$ such that $\mathbb{E}||\epsilon_{t,\theta}||^{2+\delta} \leq c_\theta$ a.s. for all $t \in \mathbb{N}, \theta \in \Theta$.

M.4. $X_{t,\theta}$ is a VAR(1) process, that is, for all $\theta \in \Theta$,

$$X_{t,\theta} = \Gamma_\theta X_{t-1,\theta} + \epsilon_{t,\theta}$$

for $t \geq 1$ and $X_{0,\theta} = 0$.

These assumptions ensure that $X_{t,\theta}$ is a VAR(1) process started at 0 with model parameters given by the index θ. Assumptions M.1 and M.4 imply that $X_{t,\theta}$ is adapted to $\mathcal{F}_{t,\theta}$.

For future reference let us define the following set of $d \times d$ matrices. For a given $\delta > 0$, let $\mathcal{J}_d(\delta) \subset \mathbb{C}^{d\times d}$ be the set of upper triangular matrices such that every $J \in \mathcal{J}_d(\delta)$ can be decomposed as $J = D+N$ with D diagonal such that $|D_{11}| \geq |D_{22}| \geq \cdots \geq |D_{dd}|$ and N equal to 0 everywhere except on the super-diagonal where it satisfies $N_{i,i+1} \in \{0, 1\}$ if $|D_{ii}| \leq \delta$ and 0 otherwise for $i = 1, \ldots, d-1$. In other words, every $J \in \mathcal{J}_d(\delta)$ can be written as a block diagonal matrix where the upper left block is diagonal and contains all eigenvalues greater than δ while the lower right block can have ones on the super-diagonal and has eigenvalues less than δ. We call the matrices in $\mathcal{J}_d(\delta)$ *Jordan-like*.

Remark 1. For any $\theta \in \Theta$, there exist matrices $F_\theta \in \mathbb{C}^{d\times d}$ and J_θ such that J_θ is a Jordan matrix and $\Gamma_\theta = F_\theta J_\theta F_\theta^{-1}$. Up to reordering of the eigenvalues, the matrix J_θ is unique and satisfies $J_\theta \in \mathcal{J}_d(|\lambda_{\theta,1}|)$. It is called the *Jordan canonical form* of Γ_θ.

2.10 Asymptotic Properties

The key building blocks for inference are the two covariance matrices

$$S_{XX} = \frac{1}{n}\sum_{t=1}^{n} X_{t-1,\theta}X_{t-1,\theta}^T, \quad S_{X\epsilon} = \frac{1}{n}\sum_{t=1}^{n} X_{t-1,\theta}\epsilon_{t,\theta}^T. \tag{2.10.1}$$

Obviously, S_{XX} and $S_{X\epsilon}$ are families of stochastic processes depending on n and θ, which we suppress to avoid cluttering up the notation, but the dependence should be kept in mind. We first need to determine what happens to S_{XX} and $S_{X\epsilon}$ when n goes to infinity and for varying θ. We cannot hope to say anything uniformly without further assumptions on Θ. The following assumptions are sufficiently general to cover a wide range of behaviours while still allowing for uniform asymptotic results.

Assumption U. Θ *satisfies the following:*

U.1. $\sup_{\theta\in\Theta} c_\theta < \infty$.

U.2. $\sup_{\theta\in\Theta}\{\sigma_{max}(\Sigma_\theta) + \sigma_{min}(\Sigma_\theta)^{-1}\} < \infty$.

U.3. There exists $\alpha \in (0,1)$ *small so that with* $r_\alpha = (1-\alpha)(2-\alpha)/\alpha$ *(see Fig. 2.10.1)*

$$\sup_{\theta\in\Theta}\left\{\max_{1\le i\le N_\theta}\left[\max\left(\frac{||\lambda_{\theta,i}|(1-\lambda_{\theta,i})|}{r_\alpha(1-|\lambda_{\theta,i}|)},\ |\lambda_{\theta,i}|\right)\right]\right\} \le 1.$$

U.4. There exists $F_\theta \in \mathbb{C}^{d\times d}$ *and* $J_\theta \in \mathcal{J}_d(1-\alpha)$ *such that* $F_\theta^{-1}\Gamma_\theta F_\theta = J_\theta$ *and*

$$\sup_{\theta\in\Theta}\left\{\sigma_{max}\left(F_\theta\right) + \sigma_{min}\left(F_\theta\right)^{-1} + \sigma_{max}\left(J_\theta\right)\right\} < \infty.$$

Assumption U.1 is a moment condition on the error process which is needed for some of the triangular array martingale difference limit results. It is implied by the other conditions if the errors are i.i.d. Gaussian. Assumption U.2 states that $||\Sigma_\theta||$ and $||\Sigma_\theta^{-1}||$ are uniformly bounded for any matrix norm. In particular, Σ_θ is of full rank. This is a natural condition when considering uniform convergence. The important assumptions are U.3 and U.4. Both assumptions have clear interpretations and are sufficient if we want to limit our attention to processes that are at most integrated of order 1 and without seasonal cointegration. In particular, to avoid higher orders of integration, we must restrict all eigenvalues to have magnitude less than 1 (ensured by Assumption U.3, see Fig. 2.10.1). For eigenvalues with magnitude 1, the corresponding Jordan block must be scalar [Archontakis, 1998] (ensured by Assumption U.4). Note that the matrices J_θ in Assumption U.4 are not required to be Jordan matrices so that the assumption allows, for example, for matrices of the form

$$\Gamma = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda' \end{pmatrix}$$

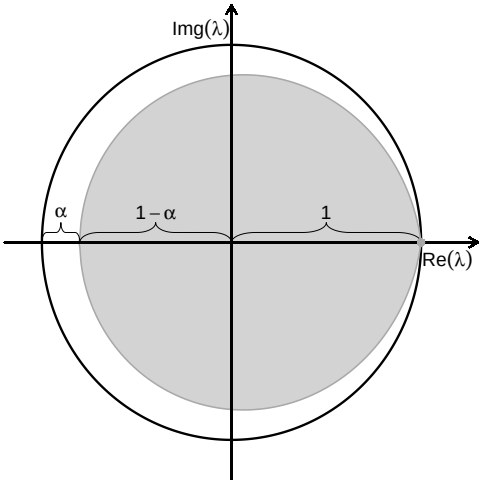

Figure 2.10.1: The region of the complex plane of allowed eigenvalues given by Assumption U.3. The gray area includes all the eigenvalues allowed in the current setting. The black circle enclosing the grey area is the unit circle. The smallest allowed real eigenvalue is $\alpha - 1$ and the only eigenvalue with magnitude 1 is real and equal to 1.

for $\lambda, \lambda' \in \mathbb{R}$ arbitrarily close together as long as $|\lambda|, |\lambda'| \leq 1 - \alpha$.

To avoid seasonal cointegration, eigenvalues with magnitude 1 are restricted to be exactly equal to 1. Specifically, as the magnitude of an eigenvalue approaches one, the eigenvalue itself approaches 1. See Fig. 2.10.1 for the region of the complex plane satisfying Assumption U.3 for some small $\alpha \in (0, 1)$. Since α can be chosen freely, the parameter space can be made arbitrarily close to the unit circle.

To state our main result, define $M_i = \sum_{j \leq i} m_j$ and write $i_k = \min\{i \geq 1 | M_i \geq k\}$. As in the univariate case [Mikusheva, 2007], we unify the range of asymptotics with an Ornstein-Uhlenbeck process. For any θ, we let $C_n(\theta)$ be the $d \times d$ diagonal matrix whose i'th diagonal block is $n \log(|\lambda_i|) I_{m_i}$ with the convention that $\log(0) = -\infty$. For convenience, we sometimes suppress the dependence on θ and n and just write C. In what follows, uniform convergence of random matrices means uniform convergence of the vectorization of these matrices so that Definition 1 applies directly.

Theorem 1 (Uniform convergence of covariance matrices)**.** *Under Assumptions M and U and after possibly enlarging* $(\Omega, \mathcal{F}, \mathbb{P})$*, there exists a standard d-dimensional Brownian motion,* $(W_t)_{t \in [0,1]}$*, and a family of processes,* $(J_{t,C})_{t \in [0,1], n \in \mathbb{N}, \theta \in \Theta}$*, with*

$$J_{t,C} = \int_0^t e^{(t-s)C} F_\theta \Sigma^{\frac{1}{2}} dW_s, \quad J_{0,C} = 0, \tag{2.10.2}$$

such that the following approximations hold for $n \to \infty$

$$H^{-\frac{1}{2}} F_\theta S_{XX} F_\theta^T H^{-\frac{1}{2}} \to_w G^{-\frac{1}{2}} \int_0^1 J_{t,C} J_{t,C}^T dt G^{-\frac{1}{2}}, \tag{2.10.3}$$

$$\sqrt{n}H^{-\frac{1}{2}}F_\theta S_{X\epsilon} \to_w G^{-\frac{1}{2}} \int_0^1 J_{t,C} dW_t^T \Sigma^{\frac{1}{2}}, \tag{2.10.4}$$

uniformly over Θ *where the covariance matrices are defined in* (2.10.1) *and the normalizing matrices are given by*

$$H = F_\theta \mathbb{E}\left(\frac{1}{n}\sum_{t=1}^n X_{t-1,\theta}X_{t-1,\theta}^T\right)F_\theta^T, \quad G = F_\theta \mathbb{E}\left(\int_0^1 J_{t,C}J_{t,C}^T dt\right)F_\theta^T. \tag{2.10.5}$$

We prove the uniform results for the special case of F_θ being the identity, i.e., under the assumption that $\Gamma \in \mathcal{J}_d(1-\delta)$. Technically, this allows for complex-valued Γ, thus, the proofs are more general than the real-valued case. Also, it is not hard to generalize to any F_θ fulfilling Assumption U.4. Indeed, for $X_{t,\theta}$ generated by $\Gamma \in \mathbb{R}^{d\times d}$, there exist $F \in \mathbb{C}^{d\times d}$ and $J \in J_d(1-\delta)$ such that F is invertible with $F^{-1}JF = \Gamma$. The transformed process $\tilde{X}_{t,\theta} = FX_{t,\theta}$ is then of the required form with parameters $\tilde{\theta} = (J, F\Sigma F^T, ||F||^{2+\delta}c)$. Assuming that F is uniformly invertible and bounded in norm then ensures that $\tilde{\theta}$ satisfies Assumption U.

We prove Theorem 1 in several steps. The main idea is to split Θ into overlapping regions and prove that Theorem 1 holds in each region. Consider

$$R_{n,0} := \left\{\theta \in \Theta : |\lambda_1| \leq 1 - \frac{\log n}{n}\right\}, \quad R_{n,d} := \left\{\theta \in \Theta : |\lambda_N| \geq 1 - n^{-\eta}\right\},$$

where $\eta \in (0,1)$ is to be specified later. The two regions correspond to the stationary and local-to-unity (non-stationary) regimes, respectively, in the univariate case. Different asymptotics arise depending on the region and, in particular, on how fast the eigenvalues converge to unity. Throughout the rest of this section we assume that Assumptions M and U hold with $F_\theta = I$.

2.10.1 Non-stationary asymptotics

In this section we consider sequences of parameters in the non-stationary region $R_{n,d}$. For simplicity we assume throughout this subsection that $\epsilon_{t,\theta}$ is i.i.d. Gaussian with mean 0 and covariance Σ_θ. In Appendix 2.H it is argued why this is not a restriction. Indeed, all the relevant sample moments can be approximated by Gaussian counterparts.

Let $(W_t)_{t\in[0,1]}$ be a standard d-dimensional Brownian motion. Since $\Sigma^{\frac{1}{2}}(W_{t/n} - W_{(t-1)/n}) \overset{\mathcal{L}}{=} \epsilon_{t,\theta}/\sqrt{n}$ for all $n \in \mathbb{N}$, $0 \leq t \leq n$ and $\theta \in \Theta$, we get

$$H^{-\frac{1}{2}}S_{X\epsilon} \overset{\mathcal{L}}{=} \int_0^1 \int_0^t f(t,s,n,\theta) dW_s dW_t^T \Sigma^{\frac{1}{2}}, \tag{2.10.6}$$

$$H^{-\frac{1}{2}}S_{XX}H^{-\frac{1}{2}} \overset{\mathcal{L}}{=} \int_0^1 \left(\int_0^t f(t,s,n,\theta)dW_s\right)\left(\int_0^t f(t,s,n,\theta)dW_s\right)^T dt, \tag{2.10.7}$$

where H is defined in (2.10.5) and $f(t,s,n,\theta) = \sqrt{n}H^{-\frac{1}{2}}\Gamma^{\lfloor nt\rfloor - \lfloor ns\rfloor - 1}\Sigma^{\frac{1}{2}}\mathbf{1}\{s \leq \lfloor nt\rfloor/n\}$. We then see that the following Lemma is a direct consequence of Lemma 5 in Appendix 2.E.

Lemma 1. *For the covariance matrices* (2.10.1) *and* H, G *and* $J_{C,t}$ *given in* (2.10.5) *and* (2.10.2)*, the following hold*

$$\lim_{n\to\infty} \sup_{\theta\in R_{n,d}} d_{BL}\left(H^{-\frac{1}{2}} S_{XX} H^{-\frac{1}{2}}, G^{-\frac{1}{2}} \int_0^1 J_{C,t} J_{C,t}^T dt G^{-\frac{1}{2}}\right) = 0, \tag{2.10.8}$$

$$\lim_{n\to\infty} \sup_{\theta\in R_{n,d}} d_{BL}\left(\sqrt{n} H^{-\frac{1}{2}} S_{X\epsilon}, G^{-\frac{1}{2}} \int_0^1 J_{C,t} dW_t^T \Sigma^{\frac{1}{2}}\right) = 0. \tag{2.10.9}$$

2.10.2 Stationary asymptotics

In this section we consider sequences of parameters in the stationary region $R_{n,0}$. We first show that the classical asymptotic theory for stationary VAR(1) processes applies. Since $R_{n,0}$ and $R_{n,d}$ overlap, we then prove that the right hand sides of (2.10.3) and (2.10.4) converge to the standard stationary limiting distributions for the diagonal entries C_{ii} going to $-\infty$.

The standard stationary theory in multiple dimensions mimics the univariate case. We follow the same strategy as in Phillips and Magdalinos [2007], but allowing for multiple dimensions and a family of error processes $\epsilon_{t,\theta}$. In this regime, we find that, when properly normalized, S_{XX} converges in probability to the identity matrix and $\mathrm{vec}(S_{Xe})$ converges in distribution to a d^2-dimensional standard Gaussian.

Lemma 2. *Let* $V \sim \mathcal{N}(0, I_{d^2})$*. For all* $\epsilon > 0$ *and* $s \in [0,1]$*,*

$$\lim_{n\to\infty} \sup_{\theta\in R_{n,0}} \mathbb{P}\left(\left\| \frac{1}{n} H^{-\frac{1}{2}} \left(\sum_{t=1}^{\lfloor ns \rfloor} X_{t-1,\theta} X_{t-1,\theta}^T\right) H^{-\frac{1}{2}} - sI \right\| > \epsilon\right) = 0 \tag{2.10.10}$$

and

$$\lim_{n\to\infty} \sup_{\theta\in R_{n,0}} d_{BL}\left(\mathrm{vec}\left(\sqrt{n} H^{-\frac{1}{2}} S_{X\epsilon} \Sigma^{-\frac{1}{2}}\right), V\right) = 0. \tag{2.10.11}$$

For the special case $s = 1$, equation (2.10.10) shows that S_{XX} converges in probability to the identity matrix. By Lemma 2 and Proposition 8 in the supplementary material of Lundborg et al. [2022], the proof of Theorem 1 in the stationary regime is complete if we can show that, for any $(\theta_n \in R_{n,0})_{n\in\mathbb{N}}$,

$$G^{-\frac{1}{2}} \int_0^1 J_{t,C} J_{t,C}^T dt G^{-\frac{1}{2}} \to_w I, \tag{2.10.12}$$

$$G^{-\frac{1}{2}} \int_0^1 J_{t,C} dW_t^T \to_w N. \tag{2.10.13}$$

We emphasize that G and C in eqs. (2.10.12)-(2.10.13) are functions of θ_n and therefore they are sequences of matrices. In particular, $C_{ii} \leq n \log(1 - \log(n)/n) \to -\infty$ for $n \to \infty$ and $1 \leq i \leq d$. Eqs. (2.10.12)-(2.10.13) are therefore a consequence of the following.

Lemma 3. *Let $(C_n)_{n\in\mathbb{N}}, (\Omega_n)_{n\in\mathbb{N}} \subset \mathbb{R}^{d\times d}$ be sequences of matrices such that C_n is diagonal, $(C_n)_{ii} \to -\infty$ for $n \to \infty$ and $1 \le i \le d$, and Ω_n is positive definite with singular values bounded from below and above uniformly over n. Let $(W_t)_{t\in[0,1]}$ be a standard d-dimensional Brownian motion and define the family of d-dimensional Ornstein-Uhlenbeck processes, $(J_{t,n})_{t\in[0,1],n\in\mathbb{N}}$, given by*

$$J_{t,n} = \int_0^t e^{(t-s)C_n}\Omega_n^{\frac{1}{2}} dW_s, \quad J_{0,n} = 0,$$

along with the normalizing matrices $G_n = \mathbb{E}\left(\int_0^1 J_{t,n}J_{t,n}^T dt\right)$. Then, for $n \to \infty$,

$$G_n^{-\frac{1}{2}} \int_0^1 J_{t,n}J_{t,n}^T dt G_n^{-\frac{1}{2}} \to_p I,$$

$$vec\left(G_n^{-\frac{1}{2}} \int_0^1 J_{t,n} dW_t^T\right) \to_w V,$$

where $V \sim \mathcal{N}(0, I_{d^2})$.

2.10.3 Mixed asymptotics

So far, all the eigenvalues were in the same regime. We have yet to explore what happens when there are eigenvalues in both regimes. We call this case the mixed regime. Define for $1 \le k \le d-1$ and some fixed $\gamma \in (0, 1-\eta)$ the sets

$$R_{n,k} = \left\{\theta \in \Theta : M_{i_k} = k, |\lambda_{i_k}| \ge 1 - n^{-\eta-\gamma}, |\lambda_{i_{k+1}}| \le 1 - n^{-\eta-\gamma}\right\}.$$

Since $1 - n^{-\eta} \le 1 - n^{-\eta-\gamma} \le 1 - \log(n)/n$, then for $\theta \in R_{n,k}$ there are at least k coordinates in the non-stationary regime and $d-k$ coordinates in the stationary regime (and some might be in both). Furthermore, for any $n \in \mathbb{N}$, $\Theta = \bigcup_{0\le k\le d} R_{n,k}$. Thus, showing that (2.10.3) and (2.10.4) hold uniformly over $R_{n,k}$ for any fixed $1 \le k \le d-1$ completes the proof of Theorem 1. This is the content of the following lemma proved in Appendix 2.E.3.

Lemma 4. *Let $1 \le k \le d-1$. We have*

$$\lim_{n\to\infty} \sup_{\theta\in R_{n,k}} d_{BL}\left(H^{-\frac{1}{2}} S_{XX} H^{-\frac{1}{2}}, G^{-\frac{1}{2}} \int_0^1 J_{C,t}J_{C,t}^T dt G^{-\frac{1}{2}}\right) = 0, \tag{2.10.14}$$

$$\lim_{n\to\infty} \sup_{\theta\in R_{n,k}} d_{BL}\left(\sqrt{n}H^{-\frac{1}{2}} S_{X\epsilon}, G^{-\frac{1}{2}} \int_0^1 J_{C,t} dW_t^T \Sigma^{\frac{1}{2}}\right) = 0. \tag{2.10.15}$$

2.11 Higher order VAR processes

The results from the VAR(1) case can be generalized to VAR(p) processes for general $p \ge 1$. First we need to specify exactly what class of models we are considering.

We extend the parameter space $\Theta \subset (\mathbb{R}^{d\times d})^p \times S_d \times \mathbb{R}_+$ and write $\theta = (\Gamma_{1,\theta}, \dots, \Gamma_{p,\theta}, \Sigma_\theta, c_\theta)$. We sometimes omit the dependence on θ in the subscript.

Assumption M(p). *$X_{t,\theta}$ and $\epsilon_{t,\theta}$ satisfy Assumptions M.1 - M.3 in addition to:*

M(p).4. $X_{t,\theta}$ is a VAR(p) process, that is, for all $\theta \in \Theta$,

$$X_{t,\theta} = \sum_{k=1}^{p} \Gamma_{k,\theta} X_{t-k,\theta} + \epsilon_{t,\theta}$$

for $t \geq 1$ and $X_{0,\theta} = \cdots = X_{1-p,\theta} = 0$.

It is well known that any VAR(p) process can reinterpreted as a VAR(1) process by writing it in its companion form. That is, one can define a matrix $\Gamma_\theta \in \mathbb{R}^{pd \times pd}$ and $\tilde{\epsilon}_{t,\theta} = (\epsilon_{t,\theta}, 0) \in \mathbb{R}^{pd}$ such that, with $Y_{t,\theta} = (X_{t,\theta}^T, \ldots, X_{t+1-p,\theta}^T)^T$,

$$Y_{t,\theta} = \Gamma_\theta Y_{t-1,\theta} + \tilde{\epsilon}_{t,\theta}.$$

Restricting the eigenstructure of Γ_θ appropriately, we can ensure that $X_{t,\theta}$ is never cointegrated of order more than 1 and never seasonally cointegrated. In particular, with a slight abuse of notation, for any $\theta \in \Theta$, we now denote the ordered eigenvalues of Γ_θ by $|\lambda_{\theta,1}| \geq |\lambda_{\theta,2}| \geq ... \geq |\lambda_{\theta,N_\theta}|$ with multiplicities $m_{\theta,1}, ..., m_{\theta,N} \in \{1, ..., pd\}$. We define M_i and i_k as above.

Assumption U(p). *Θ satisfies Assumptions U.1 - U.4 in addition to:*

U(p).5. For α as in U.3, it holds that $|\lambda_{\theta,i_{d+1}}| \leq 1 - \alpha$.

Assumption U.5 simply states that no more than d eigenvalues can be close to unity. This ensures that $\Delta X_{t,\theta}$ stays uniformly stationary over Θ or, in other words, that it is not integrated of higher orders than 1. The main difficulty in generalizing Theorem 1 is that the covariance matrix of $\tilde{\epsilon}_{t,\theta}$ is singular. The following Lemma then provides the missing piece, a proof can be found in Appendix 2.E.4.

Lemma 5. *Under Assumption U(p) the covariance $\Sigma_{Y,\theta} = \mathbb{E} Y_{p,\theta} Y_{p,\theta}^T$ is uniformly invertible, i.e., $\inf_{\theta \in \Theta} \sigma_{min}(\Sigma_{Y,\theta}) > 0$.*

Lemma 5 ensures that the normalizing matrix H is still invertible even though the noise is singular. Similar to above we define the diagonal $d \times d$ matrix C to be the one whose diagonal is given by $n \log(|\lambda_{i_k}|)$ for $k = 1, \ldots, d$. We also define the empirical covariance matrices

$$S_{YY} = \frac{1}{n} \sum_{t=1}^{n} Y_t Y_t^T, \quad S_{Y\epsilon} = \frac{1}{n} \sum_{t=1} Y_t \epsilon^T.$$

Corollary 1. *Under Assumptions M(p) and U(p) and after possibly enlarging $(\Omega, \mathcal{F}, \mathbb{P})$, there exists a standard d-dimensional Brownian motion, $(W_t)_{t \in [0,1]}$, a family of processes, $(J_{t,C})_{t \in [0,1], n \in \mathbb{N}, \theta \in \Theta}$, with*

$$J_{t,C} = \int_0^t e^{(t-s)C} F_\theta^{11} \Sigma^{\frac{1}{2}} dW_s, \quad J_{0,C} = 0,$$

where F_θ^{11} *is the upper left* $d \times d$ *block of* F_θ*, and a random matrix* $N \in \mathbb{R}^{(p-1)d \times d}$ *independent of* W_t *with* $\text{vec}(N) \sim \mathcal{N}(0, I)$ *such that the following approximations hold for* $n \to \infty$

$$H^{-\frac{1}{2}} F_\theta S_{YY} F_\theta^T H^{-\frac{1}{2}} \to_w \begin{pmatrix} G^{-\frac{1}{2}} \int_0^1 J_{t,C} J_{t,C}^T dt G^{-\frac{1}{2}} & 0 \\ 0 & I \end{pmatrix}, \tag{2.11.16}$$

$$\sqrt{n} H^{-\frac{1}{2}} F_\theta S_{Y\epsilon} \Sigma^{-\frac{1}{2}} \to_w \begin{pmatrix} G^{-\frac{1}{2}} \int_0^1 J_{t,C} dW_t^T \\ N \end{pmatrix}, \tag{2.11.17}$$

uniformly over Θ *where* $H = F_\theta \mathbb{E}(S_{YY}) F_\theta^T$ *and* $G = F_\theta^{11} \mathbb{E}\left(\int_0^1 J_{t,C} J_{t,C}^T dt \right) (F_\theta^{11})^T$.

Remark 2. Note the differences with Theorem 1. Since we are assuming that there are no more than d roots close to unity, we may split up the asymptotic expressions as in (2.11.16) and (2.11.17). As a consequence of Lemma 3 we could have equivalently phrased the result just in the way of (2.10.3) and (2.10.4), that is, without separating out the blocks. We chose this form, however, to highlight that part of the system behaves as in the stationary case.

2.12 Uniform Inference

Having established the asymptotic properties of S_{XX} and $S_{X\epsilon}$, we now seek to develop uniformly valid methods of inference. We focus on two important cases in which uniformly valid inference has so far proven challenging: predictive regression testing and coordinate confidence intervals. Inference in these settings can be hard even from a point-wise perspective since the presence of exact unit roots makes it problematic to construct test statistics with standard asymptotic distributions.

It is not trivial to conduct inference on Γ even in lieu of Theorem 1. The main problem is the presence of the nuisance parameter $C_n(\theta)$ in (2.10.3)-(2.10.4), which cannot be uniformly consistently estimated. Indeed, for a sequence $\Gamma_n = I - C/n$ where the real part of the eigenvalues of $C \in \mathbb{R}^{d \times d}$ are all strictly negative, the problem is essentially equivalent to estimating $C = n(I - \Gamma)$. But it is well known that, in this setting, Γ can only be estimated at rate $O(n^{-1})$. One way to solve this is by the use of test inversion or so-called grid bootstrap methods, which have been widely applied in the unitary case (see Hansen [1999], Mikusheva [2007] for grid bootstrap and Campbell and Yogo [2006], Phillips [2014] for an application to predictive regression). While this is fairly easy in one dimension, adapting these methods to vector autoregressive processes is prohibitive since the computational complexity quickly explodes. We now present an approach, which keeps the computational burden to a minimum. We omit the dependence on θ in the subscript of all random variables.

Consider the least squares estimator, $\hat{\Gamma}$, given by

$$\hat{\Gamma} = \frac{1}{n} \sum_{t=1}^n X_t X_{t-1}^T \left(\frac{1}{n} \sum_{t=1}^n X_{t-1} X_{t-1}^T \right)^{-1} = \Gamma + S_{X\epsilon}^T S_{XX}^{-1}.$$

It follows from Theorem 1 that $\hat{\Gamma}$ is a uniformly consistent estimator of Γ with a rate of convergence depending on the proximity of the eigenvalues of Γ to one. Indeed, since $\sqrt{n}(\hat{\Gamma}-\Gamma) = (\sqrt{n}S_{X\epsilon})^T S_{XX}^{-1}$, we find that $n(\hat{\Gamma}-\Gamma)S_{XX}(\hat{\Gamma}-\Gamma)^T = O_p(1)$ which implies that $\sqrt{n}(\hat{\Gamma}-\Gamma) = O_p(1)$. We define a uniformly consistent estimator of Σ by averaging the squared residuals, i.e., with $\hat{\epsilon}_t = X_t - \hat{\Gamma}X_{t-1}$ we define $\hat{\Sigma} = S_{\hat{\epsilon}\hat{\epsilon}}$ analogously to S_{XX} with X_{t-1} replaced by $\hat{\epsilon}_t$. Another consequence of Theorem 1 is a uniform approximation of the t^2-statistic. In particular,

$$\begin{aligned}\hat{t}_\Gamma^2 &= \operatorname{tr}\left(n\hat{\Sigma}^{-\frac{1}{2}}\left(\hat{\Gamma}-\Gamma\right)S_{XX}\left(\hat{\Gamma}-\Gamma\right)^T\hat{\Sigma}^{-\frac{1}{2}}\right)\\ &\to_w \operatorname{tr}\left(\left(\int_0^1 \hat{J}_{t,C}dW_t^T\right)^T\left(\int_0^1 \hat{J}_{t,C}\hat{J}_{t,C}^T dt\right)^{-1}\int_0^1 \hat{J}_{t,C}dW_t^T\right) := t_\Gamma^2 \quad (2.12.18)\end{aligned}$$

uniformly over Θ where $C = C_n(\theta)$ is given in Section 2.10 and $\hat{J}_{t,C}$ is defined analogously to $J_{t,C}$ but with Σ replaced by the consistent estimator $\hat{\Sigma}$. Thus, for a fixed significance level, $\alpha \in (0,1)$, a uniformly valid $100(1-\alpha)\%$ confidence region for Γ can be constructed by test-inversion. Let $q_{n,\Gamma}(1-\alpha)$ denote the $100(1-\alpha)\%$ quantile[7] of t_Γ^2 and define the confidence region

$$CR_a(\alpha; X_\theta) = \left\{\Gamma : \hat{t}_\Gamma^2 \leq q_{n,\Gamma}(1-\alpha)\right\} \quad (2.12.19)$$

where the dependence on the data $X_{0,\theta}, \ldots, X_{n,\theta}$ is made explicit in the notation. We shall sometimes omit this second argument and simply write $CR_a(\alpha)$ when this does not cause confusion. Note that the distribution of t_Γ^2 is non-standard and, therefore, computing the quantiles $q_{n,\Gamma}$ requires extensive simulations and can be quite expensive.

Another approach relies on the Gaussian approximations (Appendix 2.H). It is similar to *Andrew's Method* (see Mikusheva [2007]), and similar in spirit to grid bootstrap. It was originally suggested by Andrews [1993] but has so far only been applied in the univariate case. For a given Γ, define the VAR(1) process, $(Y_t)_{t\in\mathbb{N}}$, by

$$Y_t = \Gamma Y_{t-1} + e_t, \quad Y_0 = 0$$

where $e_t \sim \mathcal{N}(0, \hat{\Sigma})$ i.i.d. Let

$$\tilde{t}_\Gamma^2 = \operatorname{tr}\left(n\hat{\Sigma}^{-\frac{1}{2}}S_{eY}S_{YY}^{-1}S_{eY}^T\hat{\Sigma}^{-\frac{1}{2}}\right)$$

and denote by $\tilde{q}_{n,\Gamma}(1-\alpha)$ the $100(1-\alpha)\%$ quantile of $\tilde{t}_\Gamma^2$. A confidence region for Γ is then obtained by

$$CR_b(\alpha; X_\theta) = \left\{\Gamma : \tilde{t}_\Gamma^2 \leq \tilde{q}_{n,\Gamma}(1-\alpha)\right\}. \quad (2.12.20)$$

The distribution of $\tilde{t}_\Gamma^2$ is still non-standard, but the quantiles $\tilde{q}_{n,\Gamma}$ are much easier to compute by simulation.[8] The following Theorem states that both confidence regions are

[7] Note that the quantile depends on Γ (which is also made explicit in the notation) since the asymptotic distribution depends on C which is a function of Γ and n.

[8] Simulating t_Γ^2 requires numerical approximations of the stochastic integrals involved. While a small step size is generally preferred for numerical accuracy, there is also a trade off in terms of computational demands. Note, however, that samples of $\tilde{t}_\Gamma^2$ can be seen as a crude Euler approximation of t_Γ^2 with a large step size. In essence, our results state that this approximation is sufficient.

uniformly asymptotically valid over the parameter space Θ. A proof can be found in Appendix 2.F.

Theorem 2. *Fix $\alpha \in (0,1)$ and let $CR_a(\alpha; X_\theta)$ and $CR_b(\alpha; X_\theta)$ be as given in (2.12.19) and (2.12.20). Under Assumptions M and U, both are asymptotically uniformly valid over Θ in the sense that*

$$\liminf_{n\to\infty} \inf_{\theta\in\Theta} \mathbb{P}\left(\Gamma_\theta \in CR_i(\alpha; X_\theta)\right) \geq 1-\alpha$$

for $i = a, b$.

Remark 3. One might be interested in a confidence interval for a specific element of Γ. For example, testing whether X_i *Granger causes* X_j amounts to checking whether a confidence interval of Γ_{ji} contains 0. In more generality, say that we want a $100(1-\alpha)\%$ confidence interval of $g^T\Gamma f$ where $g, f \in \mathbb{R}^d$ are some arbitrary fixed vectors. Given a uniformly valid confidence region $CR(\alpha)$ of Γ, a conservative interval is then simply obtained by projection:

$$CI(\alpha) = \left(\inf_{\Gamma\in CR(\alpha)} g^T\Gamma f, \sup_{\Gamma\in CR(\alpha)} g^T\Gamma f \right).$$

Alternatively, one can first consider the t^2-statistic for the null $H_0 : \gamma^T = g^T\Gamma$, call it $\hat{t}^2_{g,\gamma}$, with uniform approximations (for a given Γ)

$$t^2_{g,\Gamma} = g^T \left(\int_0^1 \hat{J}_{t,C} dW_t^T\right)^T \left(\int_0^1 \hat{J}_{t,C}\hat{J}_{t,C}^T dt\right)^{-1} \int_0^1 \hat{J}_{t,C} dW_t^T g,$$
$$\tilde{t}^2_{g,\Gamma} = \frac{n g^T S_{eY} S_{YY}^{-1} S_{eY}^T g}{g^T \hat{\Sigma} g}.$$

In general, the distributions of $t^2_{g,\Gamma}$ and $\tilde{t}^2_{g,\Gamma}$ will unfortunately depend on the entire matrix Γ. We define $q_{n,g,\gamma}(1-\alpha) = \sup_{\Gamma: g^T\Gamma=\gamma} q_{n,g,\Gamma}(1-\alpha)$ where $q_{n,g,\Gamma}(1-\alpha)$ is the $100(1-\alpha)\%$ quantile of $t^2_{g,\Gamma}$ and define $\tilde{q}_{n,g,\gamma}$ analogously. Uniformly valid confidence regions for $g^T\Gamma$ are then given by $CR_{a,g}(\alpha) = \{\gamma : \hat{t}^2_{g,\gamma} \leq q_{n,g,\gamma}\}$ and $CR_{b,g}(\alpha) = \{\gamma : \hat{t}^2_{g,\gamma} \leq \tilde{q}_{n,g,\gamma}\}$. Finally, one can obtain uniformly valid confidence intervals for $g^T\Gamma f = \gamma^T f$ by projecting either of these confidence regions of γ.

Remark 4. In many settings it is plausible to assume that at most $k < d$ roots are in the vicinity of unity, say, $\theta \in \mathcal{R}_k = \{\theta \in \Theta : 1 - |\lambda_{i_{k+1}}| \geq \alpha\}$. This would then imply, by Lemma 7 and Lemma 8, that a simpler uniform approximation exists. Indeed, define

$$A = \int_0^1 \hat{J}_{C,t}\hat{J}_{C,t}^T dt, \quad B = \int_0^t \hat{J}_{C,t} dW_t^T,$$

and let A_k be the top left $k \times k$ block of A and B_k the first k rows of B. We then find that $\hat{t}^2_\Gamma \to \mathrm{tr}(B_k^T A_k^{-1} B_k) + \chi^2_{d-k}$ uniformly over $\mathcal{R}_k$ as n goes to infinity. This suggests a sort of hybrid approach leading to significant savings in terms of computational

requirements. Indeed, the only part that depends on Γ is the first term. Thus, if k is small, approximating $q_{n,\Gamma}(1-\alpha)$ becomes much faster. This is best illustrated by the case $k=1$ where we can write $A_1 = \int_0^1 \hat{J}_t(\lambda_1, f_1)^2 dt$ and $B_1 = \int_0^1 \hat{J}_t(\lambda_1, f_1) dW_t^T$ for

$$\hat{J}_t(\lambda_1, f_1) = \int_0^t e^{n(t-s)\log|\lambda_1|} d\hat{W}_t(f_1)$$

where $\hat{W}(f_1) = f_1 \hat{\Sigma}^{\frac{1}{2}} W_t$ and f_1 being the first row of F_θ. Now, since $\hat{W}(f_1)$ is a one dimensional brownian motion with variance $\hat{\sigma}(f_1) = f_1 \hat{\Sigma} f_1^T$, the distribution of $\text{tr}(B_1^T A_1^{-1} B_1)$ is essentially parameterized by a two-dimensional parameter $(\lambda_1, \hat{\sigma}(f_1))$. It is therefore possible to tabulate the quantiles of $\hat{t}_\Gamma^2$ for future reference. A smilar argument can be applied to the Gaussian approximation.

Remark 5. As a special case of Remark 4, we consider the VAR(p) model discussed in Section 2.11. It can be shown that inference for certain linear hypotheses can be handled with standard asymptotic theory. This all follows from the fact that the least squares estimator of the companion matrix Γ converges uniformly in distribution to a degenerate Gaussian matrix. To fix ideas, define $\Sigma_\Gamma = \lim_{n\to} F_\theta^T H^{-1} F_\theta$.[9] It follows directly from Corollary 1 that

$$\sqrt{n}(\hat{\Gamma} - \Gamma) \to_w \begin{pmatrix} \Sigma^{\frac{1}{2}} N \Sigma_\Gamma^{\frac{1}{2}} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

uniformly over Θ where $N \in \mathbb{R}^{d \times pd}$ with $\text{vec}(N) \sim \mathcal{N}(0, I)$. Additionally, Assumption U.5 ensures that $\text{rank}(\Sigma_\Gamma) \geq (p-1)d$. Thus, for any $P \in \mathbb{R}^{pd \times k}$ of rank $k \leq p(d-1)$ such that $P^T \Sigma_\Gamma P$ is of rank k, we can test the linear hypothesis $H_0 : (\Gamma_1, \ldots, \Gamma_p)P = C$ using standard results from the stationary regime. In particular, if the $kd+j$'th diagonal element of Σ_Γ is strictly positive uniformly over Θ, we can construct uniformly valid confidence intervals for $(\Gamma_k)_{ij}$ for any $i = 1, \ldots, d$ by using the standard normal quantiles scaled with the appropriate standard error. An especially relevant case is when we have reason to believe that the true data-generating process is only VAR($p-1$). Including the extra lag in our model then makes standard inference possible on $(\Gamma_1, \ldots, \Gamma_{p-1})$ (see also Section 2.12.2).

2.12.1 Predictive Regression

One application is robust inference in the predictive regression model. For a deeper discussion of why uniformly valid inference methods are important in this setting see Campbell and Yogo [2006], Elliott and Stock [1994], which cover the case of a univariate regressor, but the same considerations hold more generally. To fix ideas, consider

$$\Theta_P = \{\theta \in \Theta : \Gamma_{j1} = 0 \; \forall j = 1, ..., d\}.$$

[9] This is a well-defined positive semidefinite matrix since $\liminf_{n\to\infty} \inf_\theta \sigma_{min}(H) > 0$ and H is a monotonically increasing sequence of matrices in the sense that $H_n - H_{n-1}$ is positive semidefinite for all n.

If X_t is a VAR(1)-process satisfying Assumption M parameterized by $\theta \in \Theta_P$, i.e., the first coordinate does not affect the others, then we can split $X_t = (Y_t, \tilde{X}_t^T)^T$ and $\epsilon_t = (\rho_t, \tilde{\epsilon}_t^T)^T$ into their first coordinate and their last $d-1$ coordinates such that

$$\begin{aligned} Y_t &= \gamma^T \tilde{X}_{t-1} + \rho_t, \\ \tilde{X}_t &= \tilde{\Gamma} \tilde{X}_{t-1} + \tilde{\epsilon}_t, \end{aligned}$$

where $\gamma^T = (\Gamma_{1j})_{2 \le j \le d}$ and $\tilde{\Gamma} = (\Gamma_{ij})_{2 \le i,j \le d}$. The parameter of interest is γ. The standard approach is to compute the least squares estimator $\hat{\gamma}$ and base inference on the t^2-statistic. Unfortunately, we encounter the same issues as described above. To see this, let $\Sigma_Y = \Sigma_{11}$, $\Sigma_X = (\Sigma_{i,j})_{2 \le i,j \le d}$, $\Sigma_{YX} = (\Sigma_{1,j})_{2 \le j \le d}$, and $\Sigma_{XY} = \Sigma_{YX}^T$ and define $\delta = \Sigma_X^{-1} \Sigma_{XY}$. Then, adopting previous notation, Theorem 1 yields

$$\hat{t}_\gamma^2 \to_w \left(\int_0^1 \tilde{J}_{t,C} dB_{1,t} \right)^T \left(\int_0^1 \tilde{J}_{t,C} \tilde{J}_{t,C}^T dt \right)^{-1} \int_0^1 \tilde{J}_{t,C} dB_{1,t} =: t_\gamma^2$$

uniformly over Θ_P where $\tilde{J}_{t,C}$ consists of the last $d-1$ coordinates of $\hat{J}_{t,C}$ and $B_{1,t}$ is the first coordinate and $B_{2,t}$ the last $d-1$ coordinates of $W_t \hat{\Sigma}^{\frac{1}{2}}$. Since $B'_{1,t} = (\Sigma_Y - \delta^T \Sigma_X \delta)^{-\frac{1}{2}} (B_{1,t} - \delta B_{2,t})$ is a standard $(d-1)$-dimensional Brownian motion independent of $B_{2,t}$ satisfying $B_{1,t} = \delta B_{2,t} + (\Sigma_Y - \delta^T \Sigma_X \delta)^{\frac{1}{2}} B'_{1,t}$, we find that

$$t_\gamma^2 \overset{\mathcal{L}}{=} \left\| \left(\Sigma_Y - \delta^T \Sigma_X \delta \right)^{\frac{1}{2}} Z + Z_{\tilde{\Gamma}} \delta \right\|^2, \tag{2.12.21}$$

where $Z_{\tilde{\Gamma}} = (\int \tilde{J}_{t,C} \tilde{J}_{t,C}^T dt)^{-\frac{1}{2}} \int \tilde{J}_{t,C} dB_{2,t}$, and Z is a $(d-1)$-dimensional standard normal vector independent of $Z_{\tilde{\Gamma}}$. The nuisance parameter, $C = C_n(\theta)$, is therefore also present in the distribution of t_γ^2 via $Z_{\tilde{\Gamma}}$ necessitating alternative methods of inference. Using the results of Theorem 2, we adopt the univariate Bonferroni approach of Campbell and Yogo [2006] to obtain uniformly asymptotically valid confidence intervals. Say we want to find a confidence region for γ with significance level $\alpha \in (0,1)$. For $\alpha_1, \alpha_2 \in (0,1)$ with $\alpha_1 + \alpha_2 = \alpha$, the construction proceeds as follows: First construct a $100(1-\alpha_1)\%$ confidence region for $\tilde{\Gamma}$ using, e.g., either $CR(\alpha_1) = CR_a(\alpha_1)$ or $CR(\alpha_1) = CR_b(\alpha_1)$ (suitably modified for $d-1$ dimensions). Then, for each $\tilde{\Gamma} \in CR(\alpha_1)$, let $CR_{\gamma|\tilde{\Gamma}}(\alpha_2)$ be a $100(1-\alpha_2)\%$ confidence region for γ given $\tilde{\Gamma}$. A confidence region not depending on $\tilde{\Gamma}$ and with coverage of at least $100(1-\alpha)\%$ is then obtained via a Bonferroni correction:

$$CR_\gamma(\alpha_1, \alpha_2; X_\theta) = \bigcup_{\tilde{\Gamma} \in CR(\alpha_1; X_\theta)} CR_{\gamma|\tilde{\Gamma}}(\alpha_2; X_\theta). \tag{2.12.22}$$

Let $\hat{\gamma}_{\tilde{\Gamma}}$ be the estimator obtained by regressing $Y_{\tilde{\Gamma},t} = Y_t - \hat{\Sigma}_{YX} \hat{\Sigma}_X^{-1} (\tilde{X}_t - \tilde{\Gamma} \tilde{X}_{t-1})$ on $\tilde{X}_{t-1}$ with standard error $\hat{\sigma}_Y^2 = \hat{\Sigma}_Y - \hat{\Sigma}_{YX} \hat{\Sigma}_X^{-1} \hat{\Sigma}_{XY}$. A choice for $CR_{\gamma|\tilde{\Gamma}}(\alpha_2)$ is then given by

$$CR_{\gamma|\tilde{\Gamma}}(\alpha_2; X_\theta) = \left\{ \gamma : \hat{\sigma}_Y^{-2} \hat{t}_{\gamma|\tilde{\Gamma}}^2 \le q_{d-1,1-\alpha_2} \right\}, \tag{2.12.23}$$

with $q_{d-1,1-\alpha_2}$ denoting the $1-\alpha_2$ quantile of the χ^2_{d-1} distribution and $\hat{t}^2_{\gamma|\tilde{\Gamma}}$ the usual t^2-statistic for the estimator $\hat{\gamma}_{\tilde{\Gamma}}$ evaluated at γ. A proof of the following is given in Appendix 2.F.

Lemma 6. *For fixed significance levels $\alpha_1, \alpha_2 \in (0,1)$ with $\alpha_1+\alpha_2 \in (0,1)$, let $CR_\gamma(\alpha_1, \alpha_2; X_\theta)$ be the confidence interval given in (2.12.22) with $CR(\alpha_1; X_\theta)$ having uniform asymptotic level and $CR_{\gamma|\tilde{\Gamma}}(\alpha_2; X_\theta)$ as given in (2.12.23). Then, under Assumptions M and U,*

$$\liminf_{n\to\infty} \inf_{\theta\in\Theta_P} \mathbb{P}\left(\gamma_\theta \in CR_\gamma(\alpha_1, \alpha_2; X_\theta)\right) \geq 1-\alpha_1-\alpha_2.$$

Remark 6. Lemma 6 is easy to extend to hypothesis testing. Say, e.g., that we want to test the null of no predictive information in the regressor, $H_0 : \gamma = 0$, versus the alternative $H_A : \gamma \neq 0$. This is equivalent to checking whether $0 \in CR_\gamma(\alpha_1, \alpha_2)$. Alternatively, the test $\varphi_n : (\mathbb{R}^d)^n \to \{0,1\}$ given by $\varphi_n = \mathbf{1}\left(\inf_{\tilde{\Gamma}\in CR(\alpha_1)} \hat{\sigma}_Y^{-2}\hat{t}^2_{0|\tilde{\Gamma}} \leq q_{d-1,1-\alpha_2}\right)$ has asymptotic uniform level and does not require the explicit computation of the confidence regions $CR_{\gamma|\tilde{\Gamma}}(\alpha_2)$.

Remark 7. It is well known that the Bonferroni approach is sub-optimal in terms of power yielding confidence regions that are too conservative. Although beyond the scope of the present paper, improvements might be gained by adapting approaches similar to the ones given in Jansson and Moreira [2006] or Elliott et al. [2015].

2.12.2 Lag agumentation

Other approaches have been suggested to be robust against deviations from exact unit root assumptions. The first one is the lag-augmented VAR methodology proposed by Dolado and Lütkepohl [1996], Toda and Yamamoto [1995]. In our setup of VAR(1) processes this approach regresses X_t on X_{t-1} and the additional augmented lag X_{t-2} upon which standard inference methodology is valid. Say, for example, that we are interested in testing the hypothesis $H_0 : A\,\mathrm{vec}(\Gamma) = b$ for $A \in \mathbb{R}^{k\times d^2}$ of rank $k \leq d^2$ and $b \in \mathbb{R}^k$. Let $\hat{\Pi}_{LA} \in \mathbb{R}^{d\times 2d}$ denote the least squares estimator in the lag-augmented regression of X_t on $\bar{X}_t = (X_{t-1}^T, X_{t-2}^T)^T$. Denote $D = (I_d, 0)^T$ and define $\hat{\Gamma}_{LA} = \hat{\Pi}_{LA}D$ along with the Wald-statistic

$$\hat{t}^2_{LA,A,b} = n(A\mathrm{vec}(\hat{\Gamma}_{LA}) - b)^T\hat{\Sigma}^{-1}_{LA,A}(A\mathrm{vec}(\hat{\Gamma}_{LA}) - b), \tag{2.12.24}$$

where $\hat{\Sigma}_{LA,A} = A\hat{\Sigma}_{LA}A^T$ and $\hat{\Sigma}_{LA} = \hat{\Sigma}^{-1} \otimes \hat{\Sigma}$. Then, under the null, $\hat{t}^2_{LA,A,b}$ converges in distribution to χ^2_k uniformly over Θ allowing for construction of uniformly valid confidence intervals and tests (see Appendix 2.J). For example, a $(1-\alpha)100\%$ confidence interval for Γ_{ij} is given by

$$CI_{LA,ij}(\alpha) = \hat{\Gamma}_{LA,ij} \pm z_{1-\alpha/2}\frac{\hat{\Sigma}_{LA,ij}}{\sqrt{n}}, \tag{2.12.25}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ standard normal quantile.

Remark 8. The key ingredient that facilitates standard inference is the fact that $\sqrt{n}(\hat{\Gamma}_{LA}-\Gamma)$ converges uniformly in distribution over Θ to a family of d-dimensional Gaussians (see Lemma 1). In particular, there is no need for normalization, since all components converge at the same rate $O(\sqrt{n})$. This is contrary to the limiting behaviour of $\hat{\Gamma}$ which needs to be normalized by the matrix $H^{-\frac{1}{2}}$ since the presence of roots close to unity make certain parts of $\hat{\Gamma}$ super efficient. This also suggests some loss of efficiency when using lag augmentation, which does not come as a surprise since we are essentially overfitting the model.

2.12.3 IVX

Another approach, known as IVX, that deals specifically with the potential presence of unit roots uses endogenously constructed instrumental variables to slow down the rate of convergence of the estimator enough to ensure mixed Gaussian limiting distributions. It was first suggested by Phillips et al. [2009] and later extended in Kostakis et al. [2015], Magdalinos and Phillips [2020]. The most general framework considered so far was proposed in Magdalinos and Phillips [2020]. However, they make the crucial assumption that all roots converge to unity at the same speed. While this allows for easy construction of confidence intervals of general linear functions of Γ and simplifies the theory somewhat, this is a significant restriction. In particular, it does not yield uniform guarantees as the ones discussed in this paper. This excludes, for example, the mixed regime discussed in Section 2.10.3 covering cases where parts of the process are stationary and others exhibit random walk behaviour. In this section we detail how one may achieve truly uniform results. This comes at the cost of less general confidence regions, which is essentially because we need to employ different normalizations depending on how close the different roots are to unity. This is akin to using $\hat{t}^2_\Gamma$ for inference.

The idea of IVX is simple. We achieve Gaussian asymptotics by constructing an endogenously generated instrument that lies in the stationary regime and then perform IV-regression. For some fixed $\beta \in (1/2, 1)$, we define, for each $n \in \mathbb{N}$, the instrument $(Z_t)_{t\in\mathbb{N}}$ by

$$Z_t = (1 - n^{-\beta})Z_{t-1} + \Delta X_t, \quad Z_0 = 0,$$

where we have suppressed the dependence on n in the notation. For each $n \in \mathbb{N}$, Z_t is a VAR(1) process with the error terms given by ΔX_t and the sequence of coefficients, $(1 - n^{-\beta})I$, fall inside the stationary regime. The IVX estimator is the IV estimator of regressing X_t on X_{t-1} and the instrumental variable Z_{t-1}, i.e.,

$$\hat{\Gamma}_{IV} = \sum_{t=1}^{n} X_t Z_{t-1}^T \left(\sum_{t=1}^{n} X_{t-1} Z_{t-1}^T \right)^{-1}.$$

The corresponding t^2-statistic for testing the null $H_0 : \Gamma = \Gamma_0$ is then given by

$$\hat{t}^2_{IV,\Gamma_0} = \operatorname{tr}\left(n\hat{\Sigma}^{-\frac{1}{2}} \left(\hat{\Gamma}_{IV} - \Gamma_0\right) S_{XZ} S_{ZZ}^{-1} S_{ZX} \left(\hat{\Gamma}_{IV} - \Gamma_0\right)^T \hat{\Sigma}^{-\frac{1}{2}} \right),$$

where S_{XZ} and S_{ZZ} are defined analogously to S_{XX}. It turns out that inference based on this statistic is standard. In particular, $\hat{t}^2_{IV,\Gamma}$ has the standard asymptotic $\chi^2_{d^2}$ distribution uniformly over the parameter space Θ (see Appendix 2.K for more details) and therefore, for fixed $\alpha \in (0,1)$, a confidence region with asymptotic uniform level is given by

$$CR_{IV}(\alpha) = \left\{\Gamma : \hat{t}^2_{IV,\Gamma} \leq q_{d^2,1-\alpha}\right\}.$$

Remark 9. The use of the instrumental variable Z_t simplifies inference. Indeed, the asymptotic distribution is standard and there is no need for extensive simulations. There is, however, some loss of efficiency. The estimator $\hat{\Gamma}_{IV}$ converges at rate of n^β or slower. If β is close to one or if all the roots converge to unity at rate that is slower than n^β, this will not be a problem, but in general $\hat{\Gamma}$ is a more efficient estimator of Γ.

2.13 Simulations

In this section we investigate the finite sample properties of the methods described in the preceding section. First, we consider the problem of constructing a confidence interval for Γ_{ij}. Testing whether $X_{j,t}$ Granger causes $X_{i,t}$, which in this case amounts to testing the null $H_0 : \Gamma_{ij} = 0$, is then equivalent to checking if 0 is contained in the confidence interval. Since there is nothing special about the choice of i and j we choose to focus on Γ_{11} for simplicity. The second problem we consider is that of testing $H_0 : \gamma = 0$ in the predictive regression model.

2.13.1 Confidence intervals

Throughout we fix the significance level at $\alpha = 0.05$. We compare three different ways of constructing confidence intervals for Γ_{11}. The first is lag-augmentation yielding the confidence interval $CI_{LA}(\alpha) = CI_{LA,11}(\alpha)$, equation (2.12.25). The other two methods first compute a confidence region for the entire matrix Γ and then find a confidence interval for Γ_{11} by projecting the confidence region onto the first coordinate. That is, for a given confidence region of Γ with $(1-\alpha)100\%$ coverage, $CR(\alpha)$, we obtain the projected confidence interval

$$CI(\alpha) = \left(\inf_{\Gamma \in CR(\alpha)} \Gamma_{11}, \sup_{\Gamma \in CR(\alpha)} \Gamma_{11}\right).$$

We let $CI_b(\alpha)$ (respectively $CI_{IV}(\alpha)$) be the confidence interval obtained by projecting $CR_b(\alpha)$ (respectively $CR_{IV}(\alpha)$). Of these, CI_b is by far the most costly to compute as the dimension increases. The constraint in the optimization problem is costly to evaluate due to the need for simulations to compute the critical value $\tilde{q}_{n,\Gamma}(1-\alpha)$ at each Γ. This issue can, however, be partly resolved by the use of the EAM-algorithm [Kaido et al., 2019]. See Appendix 2.I.3 for details. Furthermore, we are being quite agnostic about the spectral structure of Γ. Imposing extra assumptions on the parameter space such as limiting the number of possible roots that are allowed to be close to unity can

Table 2.13.1: Coverage and median length of C_b, C_{IV}, and CI_{LA}.

	Coverage			Median Length		
d	CI_b	CI_{IV}	CI_{LA}	CI_b	CI_{IV}	CI_{LA}
			$n = 50$			
3	0.996	0.999	0.962	0.716	0.706	0.861
4	0.999	1.000	0.971	1.090	1.075	0.948
5	0.999	0.999	0.950	1.445	1.426	1.014
			$n = 75$			
3	0.998	0.998	0.976	0.506	0.491	0.699
4	0.999	0.999	0.973	0.786	0.756	0.758
5	1.000	1.000	0.973	1.093	1.051	0.794
			$n = 100$			
3	1.000	1.000	0.975	0.404	0.386	0.597
4	0.999	0.999	0.971	0.653	0.620	0.648
5	0.996	0.997	0.970	0.910	0.862	0.681

lead to further computational gains (see Remark 4). CI_{IV} also involves an optimization problem, but the constraint is cheap to evaluate since the critical value in CR_{IV} is fixed and standard. CI_{IV} can be computed by standard solvers. We let $\beta = 0.9$ in the IVX regression.

To verify that the confidence intervals are truly uniform, we look at choices of Γ with eigenvalues in different regimes. In particular, for a fixed dimension d and sample size n, we consider $\Gamma \in \mathbb{R}^{d\times d}$ with eigenvalues $\lambda_1(\Gamma) = 1$ and $\lambda_i(\Gamma) = 1 - (1/n)^{1/(i-1)}$ for $i = 2, ..., d$. For each simulation, we draw a new set of random eigenvectors and the errors are i.i.d. Gaussian with non-diagonal covariance matrix. For a detailed explanation of the setup, see Appendix 2.I.1. The results are recorded in Table 2.13.1.

All three confidence intervals have coverage greater than 0.95 in every case. As expected, both CI_b and CI_{IV} are conservative with practically a 100% coverage. This is already apparent in 3 dimensions. Despite the loss of efficiency, however, both yield shorter intervals than CI_{LA} in 3 dimensions for all three sample sizes. This can most likely be attributed to Γ having multiple roots close to unity, implying that the lag-augmented estimator converges at a rate slower than the IVX and the LS estimators. This advantage more or less vanishes in 4 dimensions and in 5 dimensions CI_{LA} is the clear winner. Intuitively, the dimension of the confidence regions is quadratic in d and the loss suffered by projection methods therefore quickly sets in. Interestingly, this phenomenon seems less pronounced for higher sample sizes. Another key result in Table 2.13.1 is that CI_b is wider than CI_{IV} in every case. This is counter to the fact that the least squares estimator should be more efficient. One possible explanation is that Γ has roots that converge to 1 at a slower rate than $(1/n)^\beta$ limiting the loss of efficiency of the

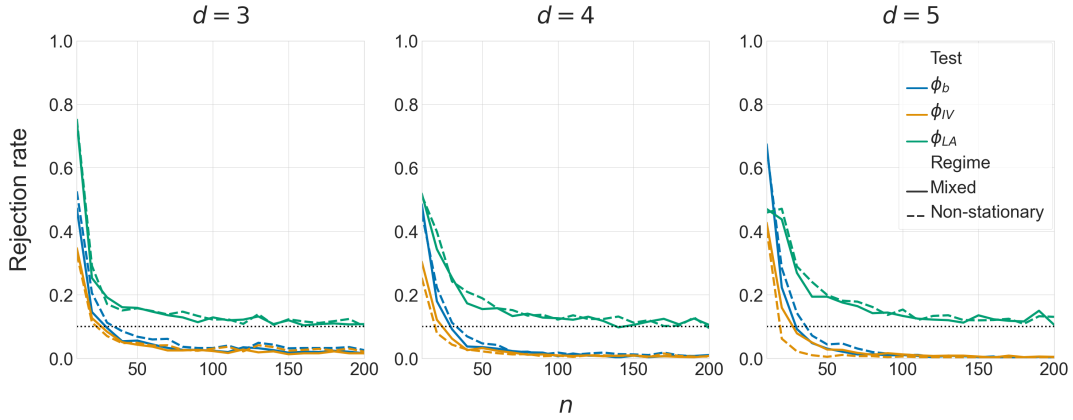

Figure 2.13.2: Rejection rates under the null $H_0 : \gamma = 0$ of the three tests at different sample sizes and dimensions and under two different regimes. The significance level is fixed at $\alpha = 0.1$ for all n and d, given by the dotted line. The rejection rate is the proportion of times the null was rejected over 1000 simulations.

IVX estimator. Another possible explanation is that finite sample behaviour of $\hat{t}^2_{IV}$ is different from the asymptotic χ^2-distribution for the sample sizes considered here, resulting in a confidence region for Γ with slightly lower coverage but still with conservative coverage when projected onto the first coordinate.

2.13.2 Predictive regression testing

Fix $\alpha = 0.1$. We compare three methods to test $H_0 : \gamma = 0$ against the alternative $H_A : \gamma \neq 0$ in the predictive regression model. The first one uses lag-augmentation and is based on the test-statistic in equation (2.12.24). We denote this test by φ_{LA}. The other two tests employ the Bonferroni strategy described in Section 2.12.1. In particular, for $\alpha_1 = 0.05$, $\alpha_2 = 0.05$, and $CR(\alpha_1)$ a confidence region for $\tilde{\Gamma}$ with uniform asymptotic level α_1, we define the test $\varphi_n = \mathbf{1}(\inf_{\tilde{\Gamma} \in CR(\alpha_1)} \hat{\sigma}_Y^{-2} \hat{t}^2_{0|\tilde{\Gamma}} \leq q_{d-1,1-\alpha_2})$. We consider $CR(\alpha_1) = CR_b(\alpha_1)$ and $CR(\alpha_1) = CR_{IV}(\alpha_1)$ denoting the corresponding tests by φ_b and φ_{IV}.[10] By Lemma 6, the tests will have uniform asymptotic level. Computing the two latter test statistics involves an optimization problem. As in the case of the projection confidence intervals, it is much costlier to compute φ_b and we employ the EAM-algorithm described in Appendix 2.I.3 as a practically feasible solution. These computational disadvantages can be reduced if one has reason to believe there are only few potential unit roots as discussed earlier.

We perform two sets of simulation experiments for the three tests. To verify that the uniform guarantees hold, we consider a sequence of $\tilde{\Gamma}$ in the mixed regime. Throughout, we let $d = 4, 5, 6$ and $\tilde{\Gamma} \in \mathbb{R}^{(d-1)\times(d-1)}$ is chosen as above. We also consider the case $\tilde{\Gamma} = I$ so that $\tilde{X}_t$ is a random walk, i.e., $\tilde{\Gamma}$ is in the non-stationary regime. First we investigate the size properties of the three tests for different sample sizes. The results

[10] Here, $CR_b(\alpha_1)$ and $CR_{IV}(\alpha_1)$ are confidence regions for $\tilde{\Gamma}$ and not the entire matrix Γ.

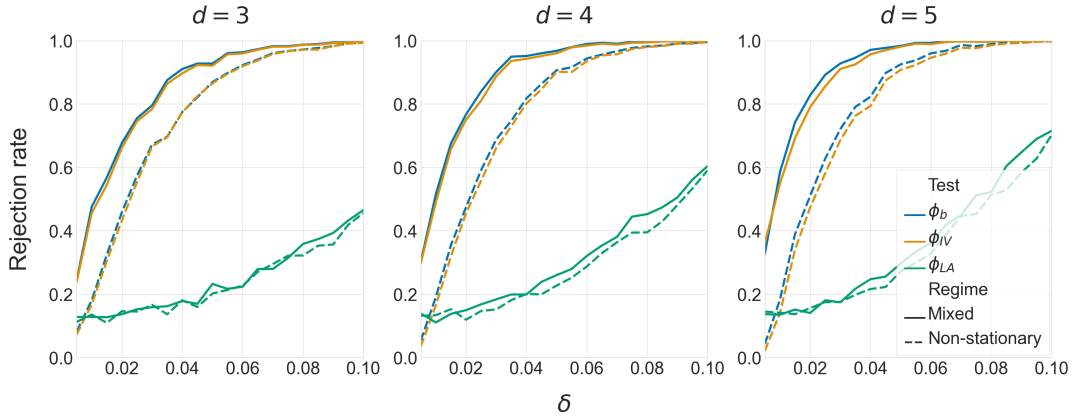

Figure 2.13.3: Rejection rates of the null $H_0 : \gamma = 0$ with $\gamma = \delta\mathbb{1}$ under different dimensions, regimes and for a sequence of alternatives. Sample size is fixed at $n = 100$, and the significance level is $\alpha = 0.1$. The rejection rate is the proportion of times the null was rejected over 1000 simulations. The dimension refers to the dimension of $\tilde{X}_t$.

are depicted in Figure 2.13.2. Evidently, φ_b and φ_{IV} quickly achieve a rejection rate well below the significance level for all three dimensions and in both regimes. This is in line with the theory since the Bonferroni corrections inherent in these tests result in tests with conservative sizes. For φ_{LA} the asymptotics take a little longer to set in. At around $n = 150$ it achieves the correct size for 3 and 4 dimensions, but convergence is slower in 5 dimensions.

To compare the power of the three tests at finite sample sizes we consider a sequence of alternatives increasingly closer to 0. In particular, we let $\gamma = \delta\mathbb{1}$ for $\delta \in \{0.005, 0.01, ..., 0.1\}$ and fix the sample size at $n = 100$. In both regimes and for all three choices of d, φ_b and φ_{IV} vastly outperform φ_{LA} correctly rejecting the null around 90% of the time in the mixed regime for $\delta = 0.04$ compared to a rejection rate of only around 18% for φ_{LA}. It looks as though φ_b is slightly better than φ_{IV} especially as the dimension increases, but the two are close overall. Interestingly, both tests seem to fare better in the mixed regime than in the stationary regime, although we would expect the LS and the IVX estimator to converge at a slower rate in the mixed regime. This might be related to the Bonferroni correction and the shape of the confidence regions in the different regimes. The power might be improved upon by using more efficient corrections than Bonferroni, see e.g. Elliott et al. [2015], Jansson and Moreira [2006]. The performance of φ_{LA} does not depend much on the regime, but it does seem to slightly improve with the size of the dimension. The latter observation also holds for the other two tests and is probably a reflection of the fact that the alternative is easier to detect in larger dimensions.

2.14 Conclusion

We proved two major uniform asymptotic results for the sample covariance matrices of VAR(1) processes with potential unit roots. First we showed that S_{XX} and $S_{X\epsilon}$ can be uniformly approximated by their Gaussian counterparts S_{YY} and $S_{Y\rho}$. This result was used to derive another uniform approximation involving integrals of Ornstein-Uhlenbeck processes. While uniform asymptotic results akin to those presented here have been given in the literature for specific sequences of Γ, this is the first time anything has been proven that is truly uniform over the parameter space Θ.

As an application of the uniform approximation results, we showed how to construct confidence regions for Γ with uniform asymptotic level. Similarly, we proved that the IVX methodology and lag augmentation also lead to uniformly valid inference if done properly.

Appendix

2.E. Proofs

Before presenting the proofs we need some auxiliary results. We assume in this section that Assumptions M and U are true with the restriction $F_\theta = I$. The first lemma collects some convergence results related to the normalizing matrix, H, defined in (2.10.5).

Lemma 1. *We have*

(a) $\liminf_{n\to\infty} \inf_{\theta\in\Theta} \sigma_{min}(H) > 0$.

(b) $\sup_{R_{n,d}} \sigma_{min}(H)^{-1} = O(n^{-\eta})$.

(c) $\sup_{R_{n,0}} \sup_{1\le k\le j\le d}(1 - |\lambda_{i_j}|)|H_{kj}| = O(1)$.

(d) $\sup_{R_{n,0}} ||\Gamma^t|| \le C(1 - \log(n)/n)^t$ *for a constant* $C \ge 0$ *not depending on* θ *or* n.

(e) $\sup_{R_{n,0}} ||\sum_{t=0}^{n-2} \Gamma^t \Sigma (\Gamma^t)^T|| = O(n/\log n)$

(f) $\sup_{R_{n,0}} ||\sum_{t=0}^{n-2} t\Gamma^t \Sigma (\Gamma^t)^T|| = O(n/\log n)$

Proof. We start with the proof of (a). We have

$$H = \frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{t-1} \Gamma^{t-1-s}\Sigma\left(\Gamma^{t-1-s}\right)^T \ge \frac{1}{n}\sum_{t=1}^{n-1}\Sigma \tag{2.E.1}$$

since every matrix in the summand is positive semidefinite. We get

$$\inf_{\theta\in\Theta} \sigma_{min}(H) \ge \inf_{\theta\in\Theta} \frac{1}{n}\sum_{t=1}^{n-1} \sigma_{min}(\Sigma) = \frac{n-1}{n} \inf_{\theta\in\Theta} \sigma_{min}(\Sigma) > \frac{n-1}{n} c$$

where $c = \inf_{\theta\in\Theta} \sigma_{min}(\Sigma) > 0$ by Assumption U.2.

For the proof of (b), note that, for any n large enough and $\theta \in R_{n,d}$, Γ is diagonal by Assumption U.4 and therefore we have from (2.E.1)

$$\begin{aligned}
\sigma_{min}(H) &\ge \frac{\sigma_{min}(\Sigma)}{n}\sum_{t=1}^{n-1}\sum_{s=0}^{t-1} |\lambda_{min}(\Gamma)|^{2s} \\
&\ge \frac{\sigma_{min}(\Sigma)}{n}\sum_{t=1}^{n-1} \frac{1-(1-n^{-\eta})^{2t}}{1-(1-n^{-\eta})^2} \\
&= \frac{\sigma_{min}(\Sigma)}{n(1-(1-n^{-\eta})^2)}\left(n - 1 - \sum_{t=1}^{n-1}(1-n^{-\eta})^{2t}\right) \\
&= \frac{\sigma_{min}(\Sigma)}{n(1-(1-n^{-\eta})^2)}\left(n - \frac{1-(1-n^{-\eta})^{2n}}{1-(1-n^{-\eta})^2}\right).
\end{aligned}$$

For n_0 large enough, we get

$$\frac{1}{n}\left(n - \frac{1-(1-n^{-\eta})^{2n}}{1-(1-n^{-\eta})^2}\right) \geq \frac{1}{2}, \quad \forall n \geq n_0$$

and, thus, with $C = 2\sup_{\theta\in\Theta} \sigma_{min}(\Sigma)^{-1} < \infty$, we have that, for all $n \geq n_0$,

$$\sup_{\theta\in R_{n,d}} \sigma_{min}(H)^{-1} \leq C\left(1 - \left(1-n^{-\eta}\right)^2\right).$$

Since $n^{\eta}(1-(1-n^{-\eta})^2) \to 2$ for $n \to \infty$ and all $\eta > 0$, this proves (b).

To prove (c), fix some $\theta \in R_{n,0}$ and note that for all $1 \leq k \leq j \leq d$

$$\left|\left(\Gamma^t\Sigma\left(\Gamma^t\right)^T\right)_{kj}\right| \leq \left\|\Gamma^t\right\|_\infty \left|\sum_{i=1}^{d}\left(\Gamma^t\Sigma\right)_{ki}\right| \leq d^2\left\|\Gamma^t\right\|_\infty \|\Sigma\|_\infty \max_{k\leq i\leq d}\left|\left(\Gamma^t\right)_{ki}\right|.$$

Now define $c_n = d^2\sup_{\theta\in R_{n,0}} \max_{t\leq n}\|\Gamma^t\|_\infty\|\Sigma\|_\infty$ and note that $\limsup_{n\to\infty} c_n < \infty$. If $|\lambda_{i_j}| \leq 1-\alpha$, since Γ is Jordan-like and $|\lambda_{i_j}| \leq 1$, we find that

$$\left(1-\left|\lambda_{i_j}\right|\right)\max_{k\leq i\leq d}\left|\left(\Gamma^t\right)_{ki}\right| \leq \binom{t}{d-1}(1-\alpha)^{t-d+1} \tag{2.E.2}$$

for $t \geq d$ in which case (c) holds. So assume that $|\lambda_{i_j}| \geq 1-\alpha$. Then it follows by assumption U.4 that

$$|H_{kj}| \leq c_n\frac{|\Sigma_{kj}|}{n}\sum_{t=1}^{n-1}\sum_{s=0}^{t-1}|\lambda_{i_j}|^s = c_n\frac{|\Sigma_{kj}|}{n(1-|\lambda_{i_j}|)}\left(n - \frac{1-|\lambda_{i_j}|^n}{1-|\lambda_{i_j}|}\right)$$

so that (c) follows from the fact that $1-|\lambda_{i_j}| \geq \log(n)/n$ for any $\theta \in R_{n,0}$.

We now prove (d). By the equivalence of the Frobenius norm and the sup norm there exists some $C \geq 0$ such that, for any $\theta \in \Theta$, it holds that

$$\|\Gamma^n\| \leq C\|\Gamma^n\|_\infty \leq C\sum_{k=1}^{d}\max_{k\leq j\leq d}\left|\left(\Gamma^n\right)_{kj}\right|.$$

Now, by eq. (2.E.2),

$$\lim_{n\to\infty}\sup_{|\lambda_{i_k}|\leq 1-\alpha}\max_{k\leq j\leq d}\left|\left(\Gamma^n\right)_{kj}\right| = 0$$

and $|\Gamma^n_{kj}| = |\lambda_{i_k}|^n\delta_{kj}$ for $|\lambda_{i_k}| > 1-\alpha$. Thus, there exists a $C \geq 0$ not depending on θ and n and such that

$$||\Gamma^t|| \leq C|\lambda_{max}(\Gamma)|^t$$

for all $\theta \in \Theta$. The result then follows from $\sup_{\theta\in R_{n,0}}|\lambda_{max}(\Gamma)| \leq 1-\log(n)/n$.

For the proof of (e), part (d) and the fact that Σ is uniformly bounded yield

$$\begin{aligned}\sup_{\theta\in R_{n,0}}\left|\left|\sum_{t=0}^{n-2}\Gamma^t\Sigma\left(\Gamma^t\right)^T\right|\right| &\leq C\sum_{t=0}^{n-2}\sup_{\theta\in R_{n,0}}||\Gamma^t||^2 \leq C\sum_{t=0}^{n-2}\left(1-\frac{\log n}{n}\right)^{2t}\\ &\leq C\sum_{t=0}^{\infty}\left(1-\frac{\log n}{n}\right)^{2t} = \frac{C}{(1-(1-\log(n)/n)^2)}\\ &= O\left(\frac{n}{\log n}\right).\end{aligned}$$

The proof of part (f) is almost the same. Indeed, by the same chain of inequalities, we find that

$$\sup_{\theta\in R_{n,0}}\left|\left|\sum_{t=0}^{n-2}t\Gamma^t\Sigma\left(\Gamma^t\right)^T\right|\right| \leq \frac{C(1-\log(n)/n)^2}{(1-(1-\log(n)/n)^2)^2} = O\left(\frac{n}{\log n}\right).$$

□

Lemma 2. *We have*

$$\lim_{n\to\infty}\sup_{\theta\in R_{n,0}}\left|\left|H^{-\frac{1}{2}}\left(\sum_{t=0}^{n-2}\Gamma^t\Sigma\left(\Gamma^t\right)^T\right)H^{-\frac{1}{2}}-I\right|\right| = 0.$$

Proof. Note that it suffices to prove that

$$\lim_{n\to\infty}\sup_{\theta\in R_{n,0}}\left|\left|\left(\sum_{t=0}^{n-2}\Gamma^t\Sigma\left(\Gamma^t\right)^T\right)^{-\frac{1}{2}}H\left(\sum_{t=0}^{n-2}\Gamma^t\Sigma\left(\Gamma^t\right)^T\right)^{-\frac{1}{2}}-I\right|\right| = 0.$$

Indeed, for any two positive definite matrices $A, B \in \mathbb{R}^{d\times d}$, we have

$$\left\|A^{-\frac{1}{2}}BA^{-\frac{1}{2}}-I\right\| \leq \left\|A^{-\frac{1}{2}}B^{\frac{1}{2}}\right\|^2\left\|B^{-\frac{1}{2}}AB^{-\frac{1}{2}}-I\right\|$$

and so, if the second term on the right hand side goes to 0, by Lemma 3 below, the term on the left hand side will also go to 0.

We have

$$\begin{aligned}H &= \frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{t-1}\Gamma^{t-1-s}\Sigma\left(\Gamma^{t-1-s}\right)^T\\ &= \frac{1}{n}\sum_{t=0}^{n-2}(n-1-t)\Gamma^t\Sigma\left(\Gamma^t\right)^T\\ &= \frac{n-1}{n}\sum_{t=0}^{n-2}\Gamma^t\Sigma\left(\Gamma^t\right)^T-\frac{1}{n}\sum_{t=1}^{n-2}t\Gamma^t\Sigma\left(\Gamma^t\right)^T.\end{aligned}$$

and, as a result,

$$\left(\sum_{t=0}^{n-2}\Gamma^t\Sigma\left(\Gamma^t\right)^T\right)^{-\frac{1}{2}} H \left(\sum_{t=0}^{n-2}\Gamma^t\Sigma\left(\Gamma^t\right)^T\right)^{-\frac{1}{2}} = \frac{n-1}{n}I - M$$

where

$$M = \left(\sum_{t=0}^{n-2}\Gamma^t\Sigma\left(\Gamma^t\right)^T\right)^{-\frac{1}{2}} \frac{1}{n}\sum_{t=1}^{n-2} t\Gamma^t\Sigma\left(\Gamma^t\right)^T \left(\sum_{t=0}^{n-2}\Gamma^t\Sigma\left(\Gamma^t\right)^T\right)^{-\frac{1}{2}}.$$

All that is left to show is therefore that M goes to 0 uniformly over $R_{n,0}$ for n going to infinity. Since each term in the sum is positive definite, we have, by Assumption U.2,

$$\liminf_{n\to\infty}\inf_{\theta\in\Theta}\sigma_{min}\left(\sum_{t=0}^{n-2}\Gamma^t\Sigma\left(\Gamma^t\right)^T\right) \geq \liminf_{n\to\infty}\inf_{\theta\in\Theta}\sigma_{min}(\Sigma) > 0$$

so that, by equivalence of the spectral norm and the Frobenius norm,

$$\limsup_{n\to\infty}\sup_{\theta\in\Theta}\left|\left|\left(\sum_{t=0}^{n-2}\Gamma^t\Sigma\left(\Gamma^t\right)^T\right)^{-\frac{1}{2}}\right|\right| < \infty.$$

By part (f) of Lemma 1, we have

$$\sup_{\theta\in R_{n,0}}\left|\left|\frac{1}{n}\sum_{t=1}^{n-2} t\Gamma^t\Sigma\left(\Gamma^t\right)^T\right|\right| = o(1).$$

Combining these results, we obtain

$$\sup_{\theta\in R_{n,0}}||M|| = o(1).$$

□

The next lemma says that checking whether A_nB_n converges to the identity matrix is the same as checking whether $A_nB_n^2A_n$ converges to the identity matrix where A_n and B_n are positive semidefinite matrices of conforming dimension.

Lemma 3. *Let $\mathcal{I}$ be some index set and consider two families of sequences of positive semidefinite $d\times d$ matrices, $(A_{n,i})_{n\in\mathbb{N},i\in\mathcal{I}}$ and $(B_{n,i})_{n\in\mathbb{N},i\in\mathcal{I}}$. Then, if*

$$\lim_{n\to\infty}\sup_{i\in\mathcal{I}}\left|\left|A_{n,i}B_{n,i}^2A_{n,i} - I\right|\right| = 0,$$

it also holds that

$$\lim_{n\to\infty}\sup_{i\in\mathcal{I}}||A_{n,i}B_{n,i} - I|| = 0.$$

Proof. First, note that

$$\liminf_{n\to\infty} \inf_{i\in\mathcal{I}} \lambda_{min}(A_{n,i}B_{n,i}) = c > 0.$$

Since $A_{n,i}B_{n,i}$ is similar to the positive definite matrix $A_{n,i}^{\frac{1}{2}}B_{n,i}A_{n,i}^{\frac{1}{2}}$, the contrary would imply the existence of a sequence $(i_n)_{n\in\mathbb{N}} \subset \mathcal{I}$ such that $\sigma_{min}(A_{n,i_n}B_{n,i_n}) \to 0$ and, thus, $\sigma_{min}(A_{n,i_n}B_{n,i_n}^2A_{n,i_n}) \to 0$ for $n \to \infty$ which, of course, is a contradiction.

Now, let $A_{n,i}B_{n,i} = U_{n,i}P_{n,i}$ be a polar decomposition, i.e., $U_{n,i}$ is orthogonal and $P_{n,i}$ positive semidefinite. Then, since $P_{n,i} = U_{n,i}^T A_{n,i}B_{n,i}$, we have $P_{n,i}^2 = P_{n,i}^T P_{n,i} = A_{n,i}B_{n,i}^2A_{n,i}$ which implies that

$$\lim_{n\to\infty} \sup_{i\in\mathcal{I}} ||P_{n,i} - I|| = 0.$$

It therefore suffices to show that $U_{n,i}$ converges to the identity matrix uniformly over $\mathcal{I}$. Since $U_{n,i}$ is orthogonal, we get

$$||U_{n,i} - I||^2 = 2d - 2\mathrm{tr}(U_{n,i}) \leq 2d \sup_{1\leq j\leq d} |1 - \lambda_j(U_{n,i})|.$$

Let $U_{n,i} = V_{n,i}D_{n,i}V_{n,i}^*$ be an eigendecomposition with $D_{n,i}$ diagonal and $V_{n,i}$ unitary. Define the Hermitian matrix $H_{n,i} = V_{n,i}^*(P_{n,i} - I)V_{n,i}$. Denote by $d_{n,i}^j$ the j'th diagonal of $D_{n,i}$ and by $h_{n,i}^{jk}$ the jk'th element of $H_{n,i}$. Since $H_{n,i} \to 0$ uniformly over $\mathcal{I}$, for fixed $\epsilon > 0$, we can pick $n_0 \in \mathbb{N}$ such that

$$\sup_{i\in\mathcal{I}} \sup_{1\leq j,k\leq d} |h_{n,i}^{jk}| < \epsilon$$

and

$$\inf_{i\in\mathcal{I}} \lambda_{min}(A_{n,i}B_{n,i}) > \frac{c}{2}$$

for all $n \geq n_0$. We define the complex disk $D_r(x) = \{y \in \mathbb{C} : |y - x| \leq r\}$ for any $x \in \mathbb{C}$ and $r > 0$. The matrix $D_{n,i} + D_{n,i}H_{n,i}$ is simlar to $A_{n,i}B_{n,i}$ so, by the Gershgorin circle theorem, for $1 \leq k \leq d$, there exists $1 \leq j \leq d$ such that

$$\lambda_k(A_{n,i}B_{n,i}) \in D_{R_j}\left(d_{n,i}^j + d_{n,i}^j h_{n,i}^{jj}\right)$$

where $R_j = \sum_{k\neq j} |h_{n,i}^{jk}|$. Recall that $|d_{n,i}^j| = 1$. Using the fact that $\lambda_k(A_{n,i}B_{n,i}) > c/2$ is real and that

$$\sup_{i\in\mathcal{I}} \lambda_k(A_{n,i}B_{n,i})^2 \leq \sup_{i\in\mathcal{I}} \sigma_{max}\left(A_{n,i}B_{n,i}^2A_{n,i}\right) \to 1$$

for $n \to \infty$, we may therefore assume that n_0 is large enough so that

$$\sup_{i\in\mathcal{I}} |\lambda_k(A_{n,i}B_{n,i}) - 1| \leq \sup_{i\in\mathcal{I}} \left|d_{n,i}^j - \lambda_k(A_{n,i}B_{n,i})\right| + \epsilon$$

for all $n \geq n_0$ which implies

$$\sup_{i\in\mathcal{I}} \left| d_{n,i}^j - 1 \right| \leq 2 \sup_{i\in\mathcal{I}} \left| d_{n,i}^j - \lambda_j \left(A_{n,i} B_{n,i} \right) \right| + \epsilon \leq 2 \sup_{i\in\mathcal{I}} R_j + \epsilon + \epsilon \leq 2d\epsilon.$$

Thus,

$$\sup_{i\in\mathcal{I}} ||U_{n,i} - I|| \leq 4d^2\epsilon$$

for all $n \geq n_0$. This completes the proof. □

Finally, we provide a lower bound for the minimum eigenvalue of a sum of two positive semi-definite matrices of a special form.

Lemma 4. *Let $A \in \mathbb{R}^{d_1\times d_1}$ and $B \in \mathbb{R}^{d_2\times d_2}$ be positive semidefinite. Let $P \in \mathbb{R}^{d_1\times d_2}$ and, with $d = d_1 + d_2$, define the $d \times d$ block-matrix*

$$C = \begin{pmatrix} A + PBP^T & PB \\ BP^T & B \end{pmatrix}.$$

Then, C is positive definite and we have $\lambda_{min}(C) \geq \min\{\lambda_{min}(A), \lambda_{min}(B)\}(2+\|P\|_2)^{-2}$.

Proof. Since B is positive definite, we can take the Schur complement of B in C, call it S, and observe that

$$S = A + PBP^T - PBB^{-1}BP^T = A.$$

Then, since both B and its schur complement are positive definite, it follows by Sylvester's Law of Inertia, that C is also positive definite. Now, to achieve the lower bound, define the $d \times d$ block matrices

$$Q = \begin{pmatrix} I & -P \\ 0 & I \end{pmatrix}, \quad S \oplus B = \begin{pmatrix} S & 0 \\ 0 & B \end{pmatrix}$$

which satisfy the well known equality $QCQ^T = S \oplus B$. It then follows that $\lambda_{min}(C) \geq \min\{\lambda_{min}(S), \lambda_{min}(B)\}\sigma_{min}(Q^{-1})^2$. But the result then follows, since $S = A$ and

$$\sigma_{min}(Q^{-1}) = \sigma_{max}(Q)^{-1} \geq (2 + \|P\|_2)^{-1}.$$

□

2.E.1 Nonstationary asymptotics

Lemma 5. *Let $g(t, s, n, \theta) = G^{-\frac{1}{2}} e^{(t-s)C} \Sigma^{\frac{1}{2}}$. We have*

$$\lim_{n\to\infty} \sup_{\theta\in R_{n,d}} \mathbb{E} \left|\left| \int_0^1 \int_0^t f(t, s, n, \theta) - g(t, s, n, \theta) dW_s dW_t^T \right|\right|^2 = 0 \tag{2.E.3}$$

and

$$\lim_{n\to\infty} \sup_{\theta\in R_{n,d}} \mathbb{E} \left|\left| \int_0^1 \left(\int_0^t f(t, s, n, \theta) dW_s \right) \left(\int_0^t f(t, s, n, \theta) dW_s \right)^T \right.\right.$$
$$\left.\left. - \left(\int_0^t g(t, s, n, \theta) dW_s \right) \left(\int_0^t g(t, s, n, \theta) dW_s \right)^T dt \right|\right|^2 = 0. \tag{2.E.4}$$

Proof of Lemma 5. We first prove (2.E.3). Define

$$h_1(t,s,n,\theta) = \sqrt{n}H^{-\frac{1}{2}}\Gamma^{\lfloor nt\rfloor-\lfloor ns\rfloor-1}\Sigma^{\frac{1}{2}}, \quad h_2(t,s,n,\theta) = \sqrt{n}H^{-\frac{1}{2}}e^{(t-s)\tilde{C}}\Sigma^{\frac{1}{2}}$$
$$h_3(t,s,n,\theta) = \sqrt{n}H^{-\frac{1}{2}}e^{(t-s)C}\Sigma^{\frac{1}{2}},$$

where $\tilde{C} = n\log\Gamma$ is well defined for all $\theta \in R_{n,d}$. When it does not cause confusion, we shall omit the arguments of functions and simply write f, g, h_1, h_2, and h_3. By applying the Itô isometry twice we find that the expectation in (2.E.3) is equal to

$$\int_0^1\int_0^t ||f-g||^2\,dsdt \leq 4\int_0^1\int_0^t ||f-h_1||^2 + ||h_1-h_2||^2 + ||h_2-h_3||^2 + ||h_3-g||^2\,dsdt$$

where the inequality is Jensen's inequality. For the first term, for any $\theta \in R_{n,d}$,

$$\begin{aligned}\int_0^1\int_0^t ||f-h_1||^2\,dsdt &= \int_0^1\int_{\lfloor nt\rfloor/n}^t \left|\left|\sqrt{n}H^{-\frac{1}{2}}\Gamma^{\lfloor nt\rfloor-\lfloor ns\rfloor-1}\Sigma^{\frac{1}{2}}\right|\right|^2 dsdt\\ &= \left(n\int_0^1 t - \frac{\lfloor nt\rfloor}{n}dt\right)\left|\left|H^{-\frac{1}{2}}\Gamma^{-1}\Sigma^{\frac{1}{2}}\right|\right|^2\\ &\leq \left|\left|H^{-\frac{1}{2}}\Gamma^{-1}\Sigma^{\frac{1}{2}}\right|\right|^2 \leq \left|\left|\Gamma^{-1}\Sigma^{\frac{1}{2}}\right|\right|^2 \operatorname{tr}\left(H^{-1}\right).\end{aligned}$$

Here Γ^{-1} is well-defined since $\theta \in R_{n,d}$. Thus,

$$\lim_{n\to\infty}\sup_{\theta\in R_{n,d}}\int_0^1\int_0^t ||f-h_1||^2\,dsdt \leq \lim_{n\to\infty}\sup_{\theta\in R_{n,d}}\left|\left|\Gamma^{-1}\Sigma^{\frac{1}{2}}\right|\right|^2\operatorname{tr}\left(H^{-1}\right) = 0$$

where we use that $\Gamma^{-1}\Sigma^{\frac{1}{2}}$ is uniformly bounded on $R_{n,d}$ in combination with Lemma 1. For the second term, we first note that, due to Assumptions U.3 and U.4, for n large enough, we may assume that $||\Gamma - I|| < 1$ for any $\theta \in R_{n,d}$. We then get $\Gamma^{\lfloor nt\rfloor-\lfloor ns\rfloor-1} = e^{(\lfloor nt\rfloor-\lfloor ns\rfloor-1)\log\Gamma}$ whence

$$\begin{aligned}||h_1-h_2||^2 &= \left|\left|\sqrt{n}H^{-\frac{1}{2}}e^{(\lfloor nt\rfloor-\lfloor ns\rfloor-1)\log\Gamma}\left(I - e^{((t-s)n-(\lfloor nt\rfloor-\lfloor ns\rfloor-1))\log\Gamma}\right)\Sigma^{\frac{1}{2}}\right|\right|^2\\ &\leq \left|\left|\sqrt{n}H^{-\frac{1}{2}}\Gamma^{\lfloor nt\rfloor-\lfloor ns\rfloor-1}\right|\right|^2 c(t,s,n,\theta)^2\,||\log\Gamma||^2\end{aligned}$$

where

$$c(t,s,n,\theta) = ((t-s)n - (\lfloor nt\rfloor - \lfloor ns\rfloor - 1))\left|\left|\Sigma^{\frac{1}{2}}\right|\right| e^{||((t-s)n-(\lfloor nt\rfloor-\lfloor ns\rfloor-1))\log\Gamma||}$$

and the inequality follows from $||e^A - e^B|| \leq ||A-B||e^{\max\{||A||,||B||\}}$ for any $A, B \in \mathbb{C}^{d\times d}$. By assumptions U.3 and U.4, we have $\lim_{n\to\infty}\sup_{\theta\in R_{n,d}}||\log\Gamma||^2 = 0$ and

$$\limsup_{n\to\infty}\sup_{\theta\in R_{n,d}}\sup_{t\in[0,1]}\sup_{s\in[0,t]} c(t,s,n,\theta)^2 \leq c_0 < \infty.$$

We also have

$$\begin{aligned}\int_0^1 \int_0^{\lfloor nt\rfloor/n} \left|\left|\sqrt{n}H^{-\frac{1}{2}}\Gamma^{\lfloor nt\rfloor-\lfloor ns\rfloor-1}\right|\right|^2 ds &= \frac{1}{n}\sum_{t=2}^{n}\sum_{s=1}^{t-1}\left|\left|H^{-\frac{1}{2}}\Gamma^{t-1-s}\right|\right|^2 \\ &\leq \left|\left|\Sigma^{-1}\right|\right| \operatorname{tr}\left(H^{-\frac{1}{2}}\mathbb{E}\left(S_{XX}\right)H^{-\frac{1}{2}}\right) \\ &= d\left|\left|\Sigma^{-1}\right|\right|.\end{aligned}$$

As shown above

$$\limsup_{n\to\infty}\sup_{\theta\in R_{n,d}}\int_0^1\int_{\lfloor nt\rfloor/n}^{t}\left|\left|\sqrt{n}H^{-\frac{1}{2}}\Gamma^{\lfloor nt\rfloor-\lfloor ns\rfloor-1}\right|\right|^2 dsdt = 0$$

so that

$$\limsup_{n\to\infty}\sup_{\theta\in R_{n,d}}\int_0^1\int_0^{t}\left|\left|\sqrt{n}H^{-\frac{1}{2}}\Gamma^{\lfloor nt\rfloor-\lfloor ns\rfloor-1}\right|\right|^2 dsdt \leq c_1 < \infty.$$

Combining these results then yields

$$\lim_{n\to\infty}\sup_{\theta\in R_{n,d}}\int_0^1\int_0^t ||h_1-h_2||^2\,dsdt \leq c_0c_1\lim_{n\to\infty}\sup_{\theta\in R_{n,d}}||\log\Gamma||^2 = 0.$$

For the third term, note that $\tilde{C}$ and C commute. A similar argument as that applied to the second term then yields

$$\lim_{n\to\infty}\sup_{\theta\in R_{n,d}}\int_0^1\int_0^t ||h_2-h_3|| = 0.$$

Finally, for the fourth term, it suffices to show that $||\sqrt{n}H^{-\frac{1}{2}}G^{\frac{1}{2}}-I||$ converges uniformly over $R_{n,b}$ to 0 for n going to infinity. Indeed, since

$$\begin{aligned}\int_0^1\int_0^t ||h_2-g||^2 &\leq \left|\left|\sqrt{n}H^{-\frac{1}{2}}G^{\frac{1}{2}}-I\right|\right|^2\int_0^1\int_0^t\left|\left|G^{-\frac{1}{2}}e^{(t-s)C}\Sigma^{-\frac{1}{2}}\right|\right|^2 dsdt \\ &= \left|\left|\sqrt{n}H^{-\frac{1}{2}}G^{\frac{1}{2}}-I\right|\right|^2\operatorname{tr}\left(G^{-\frac{1}{2}}\mathbb{E}\left(\int_0^1 J_{t,C}J_{t,C}^T dt\right)G^{-\frac{1}{2}}\right) \\ &= d\left|\left|\sqrt{n}H^{-\frac{1}{2}}G^{\frac{1}{2}}-I\right|\right|^2,\end{aligned}$$

this would imply

$$\lim_{n\to\infty}\sup_{\theta\in R_{n,d}}\int_0^1\int_0^t ||h_3-g||^2 \leq c_1\lim_{n\to\infty}\sup_{\theta\in R_{n,d}}||\sqrt{n}H^{-\frac{1}{2}}G^{\frac{1}{2}}-I|| = 0.$$

To prove the claim, we first consider $nH^{-\frac{1}{2}}GH^{-\frac{1}{2}}$. By the Itô isometry we have

$$nH^{-\frac{1}{2}}GH^{-\frac{1}{2}} = \int_0^1\int_0^t h_3h_3^T dsdt.$$

Also, similar to above,

$$\int_0^1 \int_0^t f f^T ds dt = H^{-\frac{1}{2}} \frac{1}{n} \sum_{t=2}^{n} \sum_{s=1}^{t-1} \Gamma^{t-1-s} \Sigma \left(\Gamma^{t-1-s}\right)^T H^{-\frac{1}{2}} = I.$$

Thus,

$$\left|\left| nH^{-\frac{1}{2}} G H^{-\frac{1}{2}} - I \right|\right| \leq \int_0^1 \int_0^t \left|\left| (h_3 - f) h_3^T \right|\right| ds dt + \int_0^1 \int_0^t \left|\left| f (h_3 - f)^T \right|\right| ds dt.$$

Now, since $\limsup_{n\to\infty} \sup_{\theta \in R_{n,d}} \int_0^t ||h_3||^2 + ||f||^2 ds dt < \infty$ and, as was shown above, $\lim_{n\to\infty} \sup_{\theta \in R_{n,d}} \int_0^1 \int_0^t ||h_3 - f||^2 ds dt = 0$, Hölder's inequality yields

$$\lim_{n\to\infty} \sup_{\theta \in R_{n,d}} \left|\left| nH^{-\frac{1}{2}} G H^{-\frac{1}{2}} - I \right|\right| = 0.$$

By Lemma 3 this implies that $||\sqrt{n} H^{-\frac{1}{2}} G^{\frac{1}{2}} - I||$ converges uniformly over $R_{n,b}$ to 0 for n going to infinity.

For the proof of (2.E.4) we start with the following chain of inequalities

$$\begin{aligned}
&\left|\left| \int_0^1 \int_0^t f dW_s \left(\int_0^t f dW_s \right)^T - \int_0^t g dW_s \left(\int_0^t g dW_s \right)^T dt \right|\right| \\
\leq &\left|\left| \int_0^1 \int_0^t f dW_s \left(\int_0^t f - g dW_s \right)^T dt \right|\right| + \left|\left| \int_0^1 \int_0^t f - g dW_s \left(\int_0^t g dW_s \right)^T dt \right|\right| \\
\leq &\int_0^1 \left(\left|\left| \int_0^t f dW_s \right|\right| + \left|\left| \int_0^t g dW_s \right|\right| \right) \left|\left| \int_0^t f - g dW_s \right|\right| dt \\
\leq &\left(\int_0^1 \left(\left|\left| \int_0^t f dW_s \right|\right| + \left|\left| \int_0^t g dW_s \right|\right| \right)^2 dt \right)^{\frac{1}{2}} \left(\int_0^1 \left|\left| \int_0^t f - g dW_s \right|\right|^2 dt \right)^{\frac{1}{2}} \\
\leq &\left(2 \int_0^1 \left|\left| \int_0^t f dW_s \right|\right|^2 + \left|\left| \int_0^t g dW_s \right|\right|^2 dt \right)^{\frac{1}{2}} \left(\int_0^1 \left|\left| \int_0^t f - g dW_s \right|\right|^2 dt \right)^{\frac{1}{2}}
\end{aligned}$$

where the second to last inequality is Hölder's inequality. By the Itô isometry and Fubini's theorem we have, for any $\theta \in R_{n,d}$,

$$\mathbb{E} \left(\int_0^1 \left|\left| \int_0^t f dW_s \right|\right|^2 + \left|\left| \int_0^t g dW_s \right|\right|^2 dt \right) = \int_0^1 \int_0^t ||f||^2 + ||g||^2 ds dt = 2d$$

and

$$\mathbb{E} \left(\int_0^1 \left|\left| \int_0^t f - g dW_s \right|\right|^2 dt \right) = \int_0^1 \int_0^t ||f - g||^2 ds dt$$

so that equation 2.E.4 follows by the same argument as in the proof of equation (2.E.3). □

2.E.2 Stationary asymptotics

Before proving Lemma 2 we need two auxiliary results on the rate of convergence of $X_{t-1,n}X_{t-1,n}^T$ and $S_{X\epsilon}$ akin to Lemma 3.1 in Phillips and Magdalinos [2007].

Lemma 6. *For all $s \in [0,1]$, we have*

$$\sup_{\theta \in R_{n,0}} \left|\left| X_{\lfloor ns \rfloor,\theta} X_{\lfloor ns \rfloor,\theta}^T \right|\right| = o_p\left(\frac{n}{\sqrt{\log n}}\right) \tag{2.E.5}$$

and

$$\sup_{\theta \in R_{n,0}} \left|\left| \frac{1}{n} \sum_{t=1}^{\lfloor ns \rfloor} X_{t-1,\theta} \epsilon_{t,\theta}^T \right|\right| = o_p\left(\frac{1}{\sqrt{\log n}}\right). \tag{2.E.6}$$

Proof. We first prove (2.E.5). Since $X_{\lfloor ns \rfloor,\theta} X_{\lfloor ns \rfloor,\theta}^T$ is positive semidefinite, it suffices to show that $\sup_{\theta \in R_{n,0}} \operatorname{tr}\left(\mathbb{E}\left(X_{\lfloor ns \rfloor,\theta} X_{\lfloor ns \rfloor,\theta}^T\right)\right) = o\left(\frac{n}{\sqrt{\log n}}\right)$ for all $s \in [0,1]$. Now, fix some $s \in [0,1]$. We have, for all $\theta \in \Theta$,

$$\operatorname{tr}\left(\mathbb{E}\left(X_{\lfloor ns \rfloor,\theta} X_{\lfloor ns \rfloor,\theta}^T\right)\right) \leq \operatorname{tr}\left(\mathbb{E}\left(X_{n,\theta} X_{n,\theta}^T\right)\right) = \operatorname{tr}\left(\sum_{t=0}^{n-1} \Gamma^t \Sigma \left(\Gamma^t\right)^T\right).$$

The result then follows directly from part (e) of Lemma 1.

For the proof of (2.E.6), fix some $\theta \in \Theta$ and write

$$\begin{aligned}\mathbb{E}\left(\left(\sum_{t=1}^{n} X_{t-1,\theta}\epsilon_{t,\theta}\right)\left(\sum_{t=1}^{n} X_{t-1,\theta}\epsilon_{t,\theta}\right)^T\right) &= \operatorname{tr}(\Sigma) \sum_{t=1}^{n}\sum_{s=1}^{t-1} \Gamma^{t-1-s}\Sigma\left(\Gamma^{t-1-s}\right)^T \\ &\leq \operatorname{tr}(\Sigma)\, n \sum_{t=1}^{n} \Gamma^t \Sigma \left(\Gamma^t\right)^T\end{aligned}$$

so that another application of part (e) of Lemma 1 shows that

$$\sup_{\theta \in R_{n,0}} \mathbb{E}\left|\left| \frac{1}{n} \sum_{t=1}^{n} X_{t-1,\theta}\epsilon_{t,\theta}^T \right|\right|^2 = o_p\left(\frac{1}{\log n}\right).$$

The result then follows since $\mathbb{E}\left|\left|\frac{1}{n}\sum_{t=1}^{\lfloor ns \rfloor} X_{t-1,\theta}\epsilon_{t,\theta}^T\right|\right|^2 \leq \mathbb{E}\left|\left|\frac{1}{n}\sum_{t=1}^{n} X_{t-1,\theta}\epsilon_{t,\theta}^T\right|\right|^2$ for all $s \in [0,1]$ and $\theta \in \Theta$. □

Proof of Lemma 2. We shall first tackle the proof of (2.10.10). Fix some $s \in [0,1]$ and define $\tilde{S}_{XX} = \frac{1}{n}\sum_{t=1}^{\lfloor ns \rfloor} X_{t-1,\theta} X_{t-1,\theta}^T$ for ease of notation. From the relation $X_{t,\theta} = \Gamma X_{t-1,\theta} + \epsilon_{t,\theta}$, it follows that

$$\begin{aligned}&\Gamma X_{t-1,\theta} X_{t-1,\theta}^T \Gamma^T - X_{t-1,\theta} X_{t-1,\theta}^T - \epsilon_{t,\theta}\epsilon_{t,\theta}^T \\ &\qquad = X_{t,\theta} X_{t,\theta} - X_{t-1,\theta} X_{t-1,\theta}^T - \Gamma X_{t-1,\theta}\epsilon_{t,\theta}^T - \epsilon_{t,\theta} X_{t-1,\theta}^T \Gamma^T.\end{aligned}$$

Summing over t and dividing by n then gives $\tilde{S}_{XX} = \Gamma \tilde{S}_{XX} \Gamma^T + s\Sigma - S_n$ where

$$S_n = \frac{1}{n}\left(X_{n,\theta} X_{n,\theta}^T - \sum_{t=1}^{\lfloor ns \rfloor} \left(\epsilon_{t,\theta} \epsilon_{t,\theta}^T - \Sigma \right) - \sum_{t=1}^{\lfloor ns \rfloor} \Gamma X_{t-1,\theta} \epsilon_{t,\theta}^T - \sum_{t=1}^{\lfloor ns \rfloor} \epsilon_{t,\theta} X_{t-1,\theta}^T \Gamma^T \right).$$

We can iterate this identity to get

$$\tilde{S}_{XX} = \sum_{t=0}^{\lfloor ns \rfloor - 2} \Gamma^t \Sigma \left(\Gamma^t \right)^T + \sum_{t=0}^{\lfloor ns \rfloor - 2} \Gamma^t S_n \left(\Gamma^t \right)^T + \Gamma^{\lfloor ns \rfloor - 1} \tilde{S}_{XX} \left(\Gamma^{\lfloor ns \rfloor - 1} \right)^T.$$

Now, define

$$A_n = H^{-\frac{1}{2}} \sum_{t=0}^{\lfloor ns \rfloor - 2} \Gamma^t S_n \left(\Gamma^t \right)^T H^{-\frac{1}{2}}, \quad B_n = \Gamma^{\lfloor ns \rfloor - 1} \tilde{S}_{XX} \left(\Gamma^{\lfloor ns \rfloor - 1} \right)^T.$$

By Lemma 2, it suffices to show that $\sup_{\theta \in R_{n,0}} ||A_n|| + ||B_n|| = o_p(1)$. By Lemma 6 and Lemma 2 we see that $\sqrt{\log n} S_n$ converges uniformly to 0 over $R_{n,0}$. But then, since Σ is uniformly bounded from below over Θ, for a fixed $\epsilon > 0$, we can find $n_0 \in \mathbb{N}$ large enough so that

$$\sup_{\theta \in R_{n,0}} \mathbb{P}\left(\left|\left| \sqrt{\log n} A_n \right|\right| \geq \left|\left| H^{-\frac{1}{2}} \left(\sum_{t=0}^{\lfloor ns \rfloor - 2} \Gamma^t \Sigma \left(\Gamma^t \right)^T \right) H^{-\frac{1}{2}} \right|\right| \right) < \epsilon$$

for all $n \geq n_0$. It then follows from Lemma 2 that $\sup_{\theta \in R_{n,0}} ||A_n|| = o_p(1)$. Next, we see that $\mathbb{E}B_n = \Gamma^{\lfloor ns \rfloor - 1} \tilde{H} (\Gamma^{\lfloor ns \rfloor - 1})^T$ where $\tilde{H} = \mathbb{E}(\tilde{S}_{XX})$ and therefore $\sup_{\theta \in R_{n,0}} ||\mathbb{E}B_n|| \leq C \left(1 - \frac{\log n}{n} \right)^{2(\lfloor ns \rfloor - 1)} ||\tilde{H}||$ by part (d) of Lemma 1. Finally, from the inequality $1 - \frac{n}{\log n} \leq \frac{1}{n^{1/n}}$ and part (c) of Lemma 1 we see that $\sup_{\theta \in R_{n,0}} ||\mathbb{E}B_n|| = o(1)$. Since B_n is positive semidefinite, this implies that B_n converges in probability to 0 uniformly over $R_{n,0}$ and therefore concludes the proof of (2.10.10).

For the proof of (2.10.11), let $(\theta_n)_{n \in \mathbb{N}} \subset \Theta$ be such that $\theta_n \in R_{n,0}$ and define the array $(e_{t,n})_{t \geq 1, n \in \mathbb{N}}$ by

$$e_{t,n} = \text{vec}\left(n^{-\frac{1}{2}} H^{-\frac{1}{2}} X_{t-1,\theta_n} \epsilon_{t,\theta_n}^T \Sigma_n^{-\frac{1}{2}} \right) = n^{-\frac{1}{2}} \Sigma^{-\frac{1}{2}} \epsilon_{t,\theta_n} \otimes H^{-\frac{1}{2}} X_{t-1,\theta_n}.$$

Proving (2.10.11) is equivalent to proving $\sum_{t=1}^n e_{t,n} \to_w \mathcal{N}(0, I)$. Since X_{t-1,θ_n} is measurable wrt. $\mathcal{F}_{t-1}$, we see that $e_{t,n}$ is a martingale difference array and, by (2.10.10),

$$\sum_{t=1}^n \mathbb{E}(e_{t,n} e_{t,n}^T | \mathcal{F}_{t-1}) \to_p I$$

for $n \to \infty$. Our aim is to apply the martingale difference array CLT given in Theorem 5 which amounts to checking that, for each $\gamma > 0$, $\sum_{t=1}^n \mathbb{E}\left(||e_{t,n}||^2 \mathbf{1}\left(||e_{t,n}|| > \gamma \right) | \mathcal{F}_{t-1} \right) =$

$o_p(1)$. Now, fix some $\gamma > 0$ and note that $||e_{t,n}||^2 = ||H^{-\frac{1}{2}} X_{t-1,\theta_n}||^2 ||\epsilon_{t,\theta_n}||^2$ so that

$$\begin{aligned}
&\sum_{t=1}^{n} \mathbb{E}\left(||e_{t,n}||^2 \mathbf{1}\left(||e_{t,n}|| > \gamma\right) |\mathcal{F}_{t-1}\right) \\
\leq &\frac{1}{n} \sum_{t=1}^{n} \left|\left|H^{-\frac{1}{2}} X_{t-1,\theta_n}\right|\right|^2 \mathbb{E}\left(\left|\left|\Sigma^{-\frac{1}{2}} \epsilon_{t,n}\right|\right|^2 \mathbf{1}\left(||e_{t,n}|| > \gamma\right) |\mathcal{F}_{t-1}\right) \\
\leq &C_n \max_{1 \leq t \leq n} \left\{\mathbb{E}\left(||\epsilon_{t,n}||^2 \mathbf{1}\left(||e_{t,n}|| > \gamma\right) |\mathcal{F}_{t-1}\right)\right\}
\end{aligned}$$

where $C_n = \sup_{\theta \in R_{n,0}} \operatorname{tr}\left(H^{-\frac{1}{2}} S_{XX} H^{-\frac{1}{2}}\right) \operatorname{tr}(\Sigma) = O_p(1)$ because of (2.10.10). An application of Hölder's inequality and the Markov inequality gives us

$$\begin{aligned}
&\mathbb{E}\left(||\epsilon_{t,n}||^2 \mathbf{1}\left(||e_{t,n}|| > \gamma\right) |\mathcal{F}_{t-1}\right) \\
\leq &\mathbb{E}\left(||\epsilon_{t,\theta_n}||^{2+\delta} |\mathcal{F}_{t-1,n}\right)^{\frac{2}{2+\delta}} \mathbb{P}\left(||e_{t,n}|| > \gamma |\mathcal{F}_{t-1,n}\right)^{\frac{\delta}{2+\delta}} \\
\leq &\mathbb{E}\left(||\epsilon_{t,\theta_n}||^{2+\delta} |\mathcal{F}_{t-1,n}\right)^{\frac{2}{2+\delta}} \left(\frac{\mathbb{E}(||H^{-\frac{1}{2}} X_{t-1,\theta_n}||^2 ||\Sigma^{-\frac{1}{2}} \epsilon_{t,\theta_n}||^2 |\mathcal{F}_{t-1})}{n\gamma^2}\right)^{\frac{\delta}{2+\delta}} \\
\leq &C \operatorname{tr}\left(\frac{1}{n} H^{-\frac{1}{2}} X_{t-1,\theta_n} X_{t-1,\theta_n}^T H^{-\frac{1}{2}}\right)^{\frac{\delta}{2+\delta}}
\end{aligned}$$

where $C = d^{\frac{\delta}{2+\delta}} \sup_{\theta \in \Theta} \mathbb{E}\left(||\epsilon_{t,\theta}||^{2+\delta} |\mathcal{F}_{t-1,n}\right)^{\frac{2}{2+\delta}} < \infty$ because of Assumptions M.2 and M.3. Thus, the proof is complete if we can show that

$$\sup_{\theta \in R_{n,0}} \max_{1 \leq t \leq n} \operatorname{tr}\left(\frac{1}{n} H^{-\frac{1}{2}} X_{t-1,\theta_n} X_{t-1,\theta_n}^T H^{-\frac{1}{2}}\right)^{\frac{\delta}{2+\delta}} = o_p(1).$$

This follows from the same argument as in the proof of equation (5) [Phillips and Magdalinos, 2007]. (The multivariate case is essentially the same once (2.10.10) is established.) □

Proof of Lemma 3. Define the sequence $(c_n)_{n \in \mathbb{N}} \subset \mathbb{R}^d$ given by $c_{n,i} = e^{(C_n)_{ii}/n}$ for $1 \leq i \leq d$. By assumption, we have $c_n \to 0$ for $n \to \infty$ so we can assume without loss of generality that $\max_i c_{n,i} \leq 1$ and $\min_i c_{n,i} > 0$ and, by potentially passing to a sub sequence, that c_n is monotonically decreasing.

For each n, we can then find $k_n \in \mathbb{N}$ such that

$$1 - k_n^{-\eta} \leq \min_i e^{C_{n,ii}/k_n} \leq \max_i e^{C_{n,ii}/k_n} \leq 1 - \frac{\log k_n}{k_n}.$$

Passing to another sub sequence if necessary, we may assume that k_n is strictly increasing. Now, define sequences $(\lambda_k)_{k \in \mathbb{N}} \subset \mathbb{C}^d$ and $(\Sigma_k)_{k \in \mathbb{N}} \subset \mathbb{R}^{d \times d}$ such that $|\lambda_{k,i}| = 1 - \log(k)/k$

and $\Sigma_k = I$ for $k < k_0$ and $|\lambda_{k,i}| = e^{C_{n,ii}/k_n}$ and $\Sigma_k = \Omega_n$ for $k_n \leq k < k_{n+1}$. We then have

$$1 - k^{-\eta} \leq 1 - k_n^{-\eta} \leq \min_i |\lambda_{n,i}| \leq \max_i |\lambda_{n,i}| \leq 1 - \frac{\log k_n}{k_n} \leq 1 - \frac{\log k}{k}.$$

Define $(\theta_k)_{k\in\mathbb{N}} \subset \Theta$ by $\theta_k = (\Gamma_k, \Sigma_k, c)$ where Γ_k is the diagonal matrix whose diagonal entries are given by λ_k and $c \in \mathbb{R}_+$. The above inequalities together with the assumptions on Ω_n imply that $\theta_k \in R_{k,0} \cap R_{k,d}$ for all k. Define

$$S_k = \frac{1}{k}\sum_{t=1}^{k} X_{t-1,\theta_k} X_{t-1,\theta_k}^T, \quad T_k = \frac{1}{k}\sum_{t=1}^{k} X_{t-1,\theta_k}\epsilon_{t,\theta_k}, \quad \text{and} \quad H_k = \mathbb{E}\left(S_k\right).$$

The result follows by the triangle inequality, equations (2.E.3), (2.E.4) and Lemma 2. □

2.E.3 Mixed Asymptotics

This section is devoted to the proof of Lemma 4. Assume throughout Assumptions M and U and consider the special case where $F_\theta = I$. For ease of notation we define

$$A = \int_0^1 J_{C,t} J_{C,t}^T dt, \quad B = \int_0^1 J_{C,t} dW_t^T.$$

We hold $1 \leq k \leq d-1$ fixed throughout and start by partitioning $R_{n,k}$. Let $r = d-k \geq 1$ and define $w(j,l) = (1 - |\lambda_{i_{k+j}}|)/(1 - |\lambda_{i_{k+l}}|)$ for $1 \leq j < l \leq r$. We introduce the sets

$$U_{n,0} = \left\{\theta \in R_{n,k} : w(0,1) \leq n^{-\frac{\gamma}{r}}\right\}, \quad U_{n,r} = \left\{\theta \in R_{n,k} : w(0,r) \geq n^{-\gamma}\right\},$$

$$U_{n,j} = \left\{\theta \in R_{n,k} : w(0,j) \geq n^{-\frac{j\gamma}{r}}, w(j,j+1) \leq n^{-\frac{\gamma}{r}}\right\}$$

for $j = 1, ..., r-1$. We have $R_{n,k} = \bigcup_j U_{n,j}$ for all $n \in \mathbb{N}$. Indeed, fix n and take some $\theta \in R_{n,k}$ and define $j_0 = \min\left(\inf\left\{0 \leq j \leq r-1 : w(j,j+1) \leq n^{-\frac{\gamma}{r}}\right\}, r\right)$, where we use the convention $\inf \emptyset = \infty$. If $j_0 = 0$, then clearly $\theta \in U_{n,0} = U_{n,j_0}$. Otherwise, we find that

$$w(0,j_0) = \prod_{j=0}^{j_0-1} w(j,j+1) \geq \prod_{j=0}^{j_0-1} n^{-\frac{\gamma}{r}} = n^{-\frac{j_0\gamma}{r}}$$

so that, again, $\theta \in U_{n,j_0}$. Fix some $0 \leq j \leq r$. It therefore suffices to show that (2.10.14) and (2.10.15) hold uniformly over $U_{n,j}$. To do so, we need to split the covariance matrices and the normalizing matrix into four blocks. In particular, we write

$$H = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix}$$

where H_{11} is $(k+j)\times(k+j)$ and the other blocks of conforming dimensions. Analogously, S_{XX}, $S_{X\epsilon}$, G, A and B can be written as block matrices. Block coordinates are written in the subscript when possible and otherwise in the superscript. For example, S_{XX}^{12} and A_{12} are the top right $(k+j)\times(d-k-j)$ blocks of S_{XX} and A.

Lemma 7. *For fixed $1 \leq k \leq d-1$ and $0 \leq j \leq r$, let $N \in \mathbb{R}^{(d-k-j)\times d}$ be a random matrix on $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathrm{vec}(N) \sim \mathcal{N}(0, I)$. We have the following block-wise limits*

$$\lim_{n\to\infty} \sup_{\theta\in U_{n,j}} d_{BL}\left(H_{11}^{-\frac{1}{2}} S_{XX}^{12} H_{22}^{-\frac{1}{2}}, 0\right) = 0 \tag{2.E.7}$$

$$\lim_{n\to\infty} \sup_{\theta\in U_{n,j}} d_{BL}\left(H_{11}^{-\frac{1}{2}} S_{XX}^{11} H_{11}^{-\frac{1}{2}}, G_{11}^{-\frac{1}{2}} A_{11} G_{11}^{-\frac{1}{2}}\right) = 0 \tag{2.E.8}$$

$$\lim_{n\to\infty} \sup_{\theta\in U_{n,j}} d_{BL}\left(H_{22}^{-\frac{1}{2}} S_{XX}^{22} H_{22}^{-\frac{1}{2}}, I\right) = 0. \tag{2.E.9}$$

$$\lim_{n\to\infty} \sup_{\theta\in U_{n,j}} d_{BL}\left(\sqrt{n} H_{11}^{-\frac{1}{2}} \left(S_{X\epsilon}^{11}, S_{X\epsilon}^{12}\right), G_{11}^{-\frac{1}{2}} \left(B_{11}, B_{12}\right)\right) = 0 \tag{2.E.10}$$

$$\lim_{n\to\infty} \sup_{\theta\in U_{n,j}} d_{BL}\left(\sqrt{n} H_{22}^{-\frac{1}{2}} \left(S_{X\epsilon}^{21}, S_{X\epsilon}^{22}\right), N\right) = 0 \tag{2.E.11}$$

Proof. Fix some $\theta \in U_{n,j}$. For any $i \leq i_{k+j}$, we have $|\lambda_i| \geq |\lambda_{i_{k+j}}| \geq 1 - n^{-\eta-\gamma} n^{\frac{j\gamma}{r}} \geq 1 - n^{-\eta}$ and, for any $i \geq i_{k+j}$, $|\lambda_i| \leq |\lambda_{i_k}| \leq 1 - n^{-\eta-\gamma} \leq 1 - \frac{\log n}{n}$. Equations (2.E.8), (2.E.9), (2.E.10) and (2.E.11) then follow from the proofs in Sections 2.10.1 and 2.10.2.

For the proof of (2.E.7), note that $S_{XX}^{12} = \Gamma_{11}^n S_{XX}^{12} \Gamma_{22}^n + \sum_{t=0}^{n-1} \Gamma_{11}^t S_n \left(\Gamma_{22}^t\right)^T$, where

$$S_n = S_{\epsilon\epsilon}^{12} + \frac{1}{n}\left(X_{n,\theta} X_{n,\theta}^T\right)_{12} - \Gamma_{11} S_{X\epsilon}^{12} - \left(S_{X\epsilon}^{21}\right)^T \Gamma_{22}^T$$

and $S_{\epsilon\epsilon} = (\sum_{t=1}^n \epsilon_{t,\theta} \epsilon_{t,\theta}^T)/n$. An application of Hölder's inequality yields

$$\left|\left| H_{11}^{-\frac{1}{2}} \Gamma_{11}^n S_{XX}^{12} \left(\Gamma_{22}^n\right)^T H_{22}^{-1} \right|\right| \leq \mathrm{tr}\left(H_{11}^{-\frac{1}{2}} \Gamma_{11}^n S_{XX}^{11} \left(\Gamma_{11}^n\right)^T H_{11}^{-\frac{1}{2}}\right)^{\frac{1}{2}} \times \mathrm{tr}\left(H_{22}^{-\frac{1}{2}} \Gamma_{22}^n S_{XX}^{22} \left(\Gamma_{22}^n\right)^T H_{22}^{-\frac{1}{2}}\right)^{\frac{1}{2}}$$

and it follows from (2.E.8) and (2.E.9) along with the fact that $\sup_{\theta\in U_{n,j}} ||\Gamma_{22}^n|| = o(1)$ that

$$\sup_{\theta\in U_{n,j}} \mathrm{tr}\left(H_{11}^{-\frac{1}{2}} \Gamma_{11}^n S_{XX}^{11} \left(\Gamma_{11}^n\right)^T H_{11}^{-\frac{1}{2}}\right) = O_p(1),$$

$$\sup_{\theta\in U_{n,j}} \mathrm{tr}\left(H_{22}^{-\frac{1}{2}} \Gamma_{22}^n S_{XX}^{22} \left(\Gamma_{22}^n\right)^T H_{22}^{-\frac{1}{2}}\right) = o_p(1)$$

so that $\sup_{\theta\in U_{n,j}} ||H_{11}^{-\frac{1}{2}} \Gamma_{11}^n S_{XX}^{12} \left(\Gamma_{22}^n\right)^T H_{22}^{-1}|| = o_p(1)$. For the second term, we have, for all $\theta \in U_{n,j}$,

$$\left|\left| H_{11}^{-\frac{1}{2}} \sum_{t=1}^{n-1} \Gamma_{11}^t S_n \left(\Gamma_{22}^t\right)^T H_{22}^{-\frac{1}{2}} \right|\right| \leq C_n \left|\left| H_{22}^{-\frac{1}{2}} \right|\right| \sum_{t=0}^{n-1} \left|\left| H_{11}^{-\frac{1}{2}} \Gamma_{11}^t \Sigma_{11}^{-\frac{1}{2}} \right|\right|$$

where $C_n = \sup_{\theta \in U_{n,j}} \sup_{t \geq 1} ||S_n|| ||\Sigma_{11}^{\frac{1}{2}}|| ||\Gamma_{22}^t||$. Hölder's inequality yields

$$\sum_{t=0}^{n-1} \left|\left| H_{11}^{-\frac{1}{2}} \Gamma_{11}^t \Sigma_{11}^{-\frac{1}{2}} \right|\right| \leq \text{tr} \left(H_{11}^{-\frac{1}{2}} \sum_{t=0}^{n-1} \Gamma_{11}^t \Sigma_{11} \left(\Gamma_{11}^t\right)^T H_{11}^{-\frac{1}{2}} \right)$$

so that, by part (b) of Lemma 1 and Lemma 2,

$$\sup_{\theta \in U_{n,j}} \left|\left| H_{22}^{-\frac{1}{2}} \right|\right| \sum_{t=0}^{n-1} \left|\left| H_{11}^{-\frac{1}{2}} \Gamma_{11}^t \Sigma_{11}^{-\frac{1}{2}} \right|\right| = o_p(1).$$

Since $||\Sigma_{11}^{\frac{1}{2}}||$ and $\sup_{t \geq 1} ||\Gamma_{22}^t||$ are uniformly bouded over $U_{n,j}$, it therefore suffices to show that $\sup_{\theta \in U_{n,j}} ||S_n|| = O_p(1)$. From Lemma 2 and part (b) of Lemma 3 in the Appendix we have $\sup_{\theta \in U_{n,j}} \left|\left| S_{\epsilon\epsilon} + \frac{1}{n} X_{n,\theta} X_{n,\theta} \right|\right| = O_p(1)$ and it follows from (2.E.10) and (2.E.11) along with the fact that $H_{11}, H_{22} = O(n)$ uniformly over Θ that $\sup_{\theta \in \Theta} ||S_{X\epsilon}|| = O_p(1)$ which completes the proof. □

We now define $\tilde{H}$ as the block diagonal matrix obtained by deleting the off-diagonal blocks of H. Lemma 7 determines the limiting behaviour of $\tilde{H}^{-\frac{1}{2}} S_{XX} \tilde{H}^{-\frac{1}{2}}$. The next lemma explains why this is sufficient.

Lemma 8. *For fixed $1 \leq k \leq d-1$ and $0 \leq j \leq r$, let $\tilde{H}$ and $\tilde{G}$ be the block-diagonal matrices obtained by deleting the off-diagonal blocks of H and G, respectively. We then have*

$$\sup_{\theta \in U_{n,j}} \left\{ \left|\left| H^{-\frac{1}{2}} \tilde{H}^{\frac{1}{2}} - I \right|\right| + \left|\left| G^{-\frac{1}{2}} \tilde{G}^{\frac{1}{2}} - I \right|\right| \right\} = o(1).$$

Proof. It suffices to show that $\sup_{\theta \in U_{n,j}} ||H_{11}^{-\frac{1}{2}} H_{12} H_{22}^{-\frac{1}{2}}|| = o(1)$. To do so, we first note that, arguing as in the proof of part (b) of Lemma 1, $\sup_{\theta \in U_{n,j}} \sigma_{min}(H_{11}^{-1}) = O(1 - |\lambda_{i_{k+j}}|)$ and, consequently,

$$\sup_{\theta \in U_{n,j}} \left|\left| H_{11}^{-\frac{1}{2}} \left(1 - |\lambda_{i_{k+j}}|\right)^{-\frac{1}{2}} \right|\right| = O(1).$$

Let $\Lambda \in \mathbb{R}^{(d-k-j)\times(d-k-j)}$ be the diagonal matrix satisfying $\Lambda_{ll} = 1 - |\lambda_{i_{k+j+l}}|$. Then, by part (c) of Lemma 1, we have

$$\sup_{\theta \in U_{n,j}} \left|\left| H_{12} \Lambda^{\frac{1}{2}} \right|\right| = O\left(\left(1 - |\lambda_{i_{k+j+1}}|\right)^{-\frac{1}{2}} \right).$$

Because of the Jordan-like nature of Γ and, for any $\theta \in U_{n,j}$,

$$\begin{aligned} \sigma_{min}\left(\Lambda^{\frac{1}{2}} H_{22} \Lambda^{\frac{1}{2}} \right) &\geq \frac{\sigma_{min}(\Sigma_{22})}{n} \sum_{t=1}^{n-1} \sum_{s=0}^{t-1} \sigma_{min}\left(\Lambda^{\frac{1}{2}} \Gamma^s \right)^2 \\ &\geq \sigma_{min}\left(\Sigma_{22}\right) \min_{1 \leq l \leq d-k-j} \frac{\Lambda_{ll}}{n} \sum_{t=1}^{n-1} \sum_{s=0}^{t-1} \min\left\{ \left|\lambda_{i_{k+j+l}}\right|^{2s}, 1 \right\} \end{aligned}$$

where the second inequality follows from Assumption U.4 and the fact that $\Gamma^0 = I$. For any $1 \leq l \leq d-k-j$, it holds that

$$\begin{aligned}\frac{\Lambda_{ll}}{n}\sum_{t=1}^{n-1}\sum_{s=0}^{t-1}\left|\lambda_{i_{k+j+l}}\right|^{2s} &= \frac{1-|\lambda_{i_{k+j+l}}|}{n}\sum_{t=1}^{n-1}\sum_{s=0}^{t-1}\left|\lambda_{i_{k+j+l}}\right|^{2s} \\ &= \frac{1}{1+|\lambda_{i_{k+j+l}}|}\left(1-\frac{1-|\lambda_{i_{k+j+l}}|^{2n}}{n(1-|\lambda_{i_{k+j+l}}|^2)}\right).\end{aligned}$$

Now, since $\sup_{\theta\in U_{n,j}}|\lambda_{i_{k+j+l}}|^{2n} \to 0$ and $\inf_{\theta\in U_{n,j}} n(1-|\lambda_{i_{k+j+l}}|^2) \to \infty$ for $n\to\infty$ and $1 \leq l \leq d-k-j$, we get that

$$\sup_{\theta\in U_{n,j}} \sigma_{max}\left(\Lambda^{-\frac{1}{2}}H_{22}^{-1}\Lambda^{-\frac{1}{2}}\right) = \sup_{\theta\in U_{n,j}} \sigma_{min}\left(\Lambda^{\frac{1}{2}}H_{22}\Lambda^{\frac{1}{2}}\right)^{-1} = O(1)$$

from which it follows that $\sup_{\theta\in U_{n,j}}||\Lambda^{-\frac{1}{2}}H_{22}^{-\frac{1}{2}}|| = O(1)$. Combining all these rates yields

$$\sup_{\theta\in U_{n,j}}\left|\left|H_{11}^{-\frac{1}{2}}H_{12}H_{22}^{-\frac{1}{2}}\right|\right| = O\left(\frac{1-|\lambda_{i_{k+j}}|}{1-|\lambda_{i_{k+j+1}}|}\right)^{\frac{1}{2}} = O\left(n^{-\frac{\gamma}{2r}}\right).$$

□

With these two lemmas we can complete the proof of (2.10.14) and (2.10.15). First, for any $\theta \in U_{n,j}$, we have

$$\left|\left|H^{-\frac{1}{2}}S_{XX}H^{-\frac{1}{2}} - \tilde{H}^{-\frac{1}{2}}S_{XX}\tilde{H}^{-\frac{1}{2}}\right|\right| \leq C_n\left|\left|\tilde{H}^{-\frac{1}{2}}S_{XX}\tilde{H}^{-\frac{1}{2}}\right|\right|$$

where, by Lemma 8, $C_n = \sup_{\theta\in U_{n,j}}\left|\left|\tilde{H}^{\frac{1}{2}}H^{-\frac{1}{2}} - I\right|\right|\left(\left|\left|\tilde{H}^{\frac{1}{2}}H^{-\frac{1}{2}}\right|\right| + \sqrt{d}\right) = o(1)$. It then follows from Lemma 7 that

$$\sup_{\theta\in U_{n,j}}\left|\left|H^{-\frac{1}{2}}S_{XX}H^{-\frac{1}{2}} - \tilde{H}^{-\frac{1}{2}}S_{XX}\tilde{H}^{-\frac{1}{2}}\right|\right| = o_p(1)$$

and, similarly,

$$\begin{aligned}\sup_{\theta\in U_{n,j}}\left|\left|G^{-\frac{1}{2}}AG^{-\frac{1}{2}} - \tilde{G}^{-\frac{1}{2}}A\tilde{G}^{-\frac{1}{2}}\right|\right| &= o_p(1), \\ \sup_{\theta\in U_{n,j}}\left|\left|\sqrt{n}H^{-\frac{1}{2}}S_{X\epsilon} - \sqrt{n}\tilde{H}^{-\frac{1}{2}}S_{X\epsilon}\right|\right| &= o_p(1), \\ \sup_{\theta\in U_{n,j}}\left|\left|G^{-\frac{1}{2}}B - \tilde{G}^{-\frac{1}{2}}B\right|\right| &= o_p(1).\end{aligned}$$

Finally, arguing as in the proof of Lemma 3, we find

$$\lim_{n\to\infty} \sup_{\theta\in U_{n,j}} d_{BL}\left(G_{11}^{-\frac{1}{2}} A_{12} G_{22}^{-\frac{1}{2}}, 0\right) = 0$$
$$\lim_{n\to\infty} \sup_{\theta\in U_{n,j}} d_{BL}\left(G_{22}^{-\frac{1}{2}} A_{22} G_{22}^{-\frac{1}{2}}, I\right) = 0$$
$$\lim_{n\to\infty} \sup_{\theta\in U_{n,j}} d_{BL}\left(G_{11}^{-\frac{1}{2}} (B_{11}, B_{12}), N\right) = 0$$

so that (2.10.14) and (2.10.15) follow from Lemma 7.

2.E.4 Higher order VAR processes

Proof of Lemma 5. We assume throughout that $p \geq 2$ (since the case for $p = 1$ is already covered). For any $k = 1, \ldots, p$ and $A \in \mathbb{R}^{dp\times dp}$ we shal write $A(k)$ to denote the top-left $dk \times dk$ block and we define the matrices $\Sigma_{Y,k} = \sum_{0\leq t\leq k-1} \Gamma(k)^t \tilde{\Sigma}(k)(\Gamma(k)^t)^T$ where $\tilde{\Sigma}$ denotes the covariance matrix of $\tilde{\epsilon}_t$. We note that $\Sigma_Y = \Sigma_{Y,p}$. Now, we claim that $\Sigma_{Y,k}$ satisfies the recursive relation,

$$\Sigma_{Y,k} = \begin{pmatrix} \Sigma_{Y,k-1} + P_k \Sigma P_k^T & P_k \Sigma \\ \Sigma P_k^T & \Sigma \end{pmatrix}, \quad \Sigma_{Y,0} = \Sigma \tag{2.E.12}$$

where $P_k = (I_{d(k-1)}, 0) Q_k \in \mathbb{R}^{d(k-1)\times d}$ and $Q_k \in \mathbb{R}^{dp\times d}$ consists of the left-most $dp \times d$ block of Γ^{k-1}. One can see this by first writing

$$\Sigma_{Y,k} = \sum_{t=0}^{k-2} \Gamma(k)^t \tilde{\Sigma}(k) \left(\Gamma(k)^t\right)^T + \Gamma(k)^{k-1} \tilde{\Sigma}(k) \left(\Gamma(k)^{k-1}\right)^T.$$

Only the top-left $d \times d$ block of $\tilde{\Sigma}$ is non-zero and the bottom-left $d \times d$ block of $\Gamma(k)^t$ is 0 for all $t \leq k-2$ by the special form of the companion matrix. Thus, the first term of the sum above is given by

$$\begin{pmatrix} \Sigma_{Y,k-1} & 0 \\ 0 & 0 \end{pmatrix}.$$

The left-most $d(k-1) \times d$ block of $\Gamma(k)^{k-1}$ equals $(P_k^T, I)^T$ and therefore the second term is equal to

$$\begin{pmatrix} P_k \\ I \end{pmatrix} \Sigma \begin{pmatrix} P_k^T & I \end{pmatrix} = \begin{pmatrix} P_k \Sigma P_k^T & P_k \Sigma \\ \Sigma P_k^T & \Sigma \end{pmatrix}$$

from which the recursive relation (2.E.12) then follows.

Now to prove that Σ_Y is uniformly invertible across Θ, we shall use an induction argument. In particular, note that by Assumption U.2, the fact that Γ is uniformly bounded over Θ, and Lemma 4, $\inf_{\theta\in\Theta} \sigma_{min}(\Sigma_{Y,k-1}) > 0$ implies that the same holds for $\Sigma_{Y,k}$ for all $k = 1, \ldots, p-1$. But then the result follows if we can just show that $\inf_{\theta\in\Theta} \Sigma_{Y,0} > 0$ which of course holds under U.2 since $\Sigma_{Y,0} = \Sigma$.

□

Proof of Corollary 1. Since only d eigenvalues can be close to unity, we have $\Theta \subset \bigcup_{k=0}^{d} R_{n,k}$ and $R_{n,k} = \bigcup_{j=0}^{d-k} U_{n,j}^k$ where $R_{n,k}$ and $U_{n,j}^k$ are as in the previous section only now we make the dependence of $U_{n,j}^k$ on $R_{n,k}$ explicit by adding the superscript k. Fix some $k \in \{0,1,\dots,d\}$ and $j \in \{0,1,\dots,d-k\}$. It then suffices to show that the result holds uniformly over $U_{n,j}^k$.

We split H, S_{YY}, and $S_{Y\tilde{\epsilon}}$ into four blocks with the top left block being $(k+j)\times(k+j)$ and the other blocks of conforming dimensions. One readily sees that, since $k+j \leq d$ and $\Sigma \in \mathbb{R}^{d\times d}$ is uniformly of full rank, Lemma 1.(b) holds for H_{11}. Similarly, by Lemma 5 we find that Lemma 1.(a) and Lemma 2 hold for H_{22} (all other parts of Lemma 1 hold also without reference to the Lemma 5). Thus, the proof of Lemma 7 carries over without any modifications. Defining

$$A = \int_0^1 J_{C,t} J_{C,t}^T dt, \quad B = \int_0^1 J_{C,t} dW_t^T,$$

and $N \in \mathbb{R}^{(pd-j)\times(pd-j)}$ with $\text{vec}(N) \sim \mathcal{N}(0, I)$ independent of A and B, we therefore find that

$$\tilde{H}^{-\frac{1}{2}} F_\theta S_{YY} F_\theta^T \tilde{H}^{-\frac{1}{2}} \to_w \begin{pmatrix} G_{11}^{-\frac{1}{2}} A_{11} G_{11}^{-\frac{1}{2}} & 0 \\ 0 & I \end{pmatrix},$$

$$\sqrt{n}\tilde{H}^{-\frac{1}{2}} F_\theta S_{Y\tilde{\epsilon}} \to_w \begin{pmatrix} G_{11}^{-\frac{1}{2}} (B_{11}, B_{12}) \\ N \end{pmatrix},$$

uniformly over $U_{n,j}$. Here, A_{11} denotes the upper left $(k+j)\times(k+j)$ block of A and similarly for B_{11} and G_{11}. $\tilde{H}$ is the block diagonal matrix obtained by deleting the off-diagonal blocks of H. This is not quite the result we want. First, we replace $\tilde{H}$ with H by virtue of Lemma 8. Second, to get the right hand sides of (2.11.16) and (2.11.17), we need to show that

$$\lim_{n\to\infty} \sup_{\theta \in U_{n,j}^k} d_{BL}\left(\begin{pmatrix} G_{11}^{-\frac{1}{2}} A_{11} G_{11}^{-\frac{1}{2}} & 0 \\ 0 & I_{d-k} \end{pmatrix}, G^{-\frac{1}{2}} A G^{-\frac{1}{2}} \right) = 0$$

$$\lim_{n\to\infty} \sup_{\theta \in U_{n,j}^k} d_{BL}\left(\begin{pmatrix} G_{11}^{-\frac{1}{2}} (B_{11}, B_{12}) \\ (I_{d-k}, 0) N \end{pmatrix}, G^{-\frac{1}{2}} B \right) = 0.$$

This follows from first an application of Lemma 8 and then Lemma 3. □

2.F. Confidence Regions

This section captures some of the more technical details omitted from Section 2.12. Validity of $CR_a(\alpha)$ and $CR_b(\alpha)$ is a fairly straightforward consequence of the fact that $\hat{t}_\Gamma^2$ can be uniformly approximated by t_Γ^2 and $\tilde{t}_\Gamma^2$ both of which have continuous distributions.

Proof of Theorem 2. The result follows from Proposition 13 in the supplementary material of Lundborg et al. [2022] since $\hat{t}_\Gamma^2$ and $\tilde{t}_\Gamma^2$ both converge in distribution to t_Γ^2 uniformly over Θ and the latter is uniformly absolutely continuous wrt. Lebesgue measure. □

2.F.1 Predictive regression

Proof of Lemma 6. For each $\theta \in \Theta_P$ with γ and $\tilde{\Gamma}$ the corresponding autoregressive coefficients, define the events

$$A_\theta = \left\{\omega \in \Omega : \tilde{\Gamma} \notin CR(\alpha_1, \omega)\right\}, \quad B_\theta = \left\{\omega \in \Omega : \gamma \notin C_{\gamma|\tilde{\Gamma}}(\alpha_2, \omega)\right\}$$

where the dependence of the confidence regions on ω is made explicit in the notation. If $\omega \in \Omega$ is such that $\gamma \notin CI_\gamma(\alpha_1, \alpha_2, \omega)$, then we must either have $\omega \in A_\theta$ or $\omega \in B_\theta$ implying, by Bonferroni's inequality, $\mathbb{P}\left(\gamma \notin CI_\gamma(\alpha_1, \alpha_2)\right) \leq \mathbb{P}(A_\theta) + \mathbb{P}(B_\theta)$. It follows by assumption that $\limsup_{n\to\infty} \sup_{\theta\in\Theta_P} \mathbb{P}(A_\theta) \leq \alpha_1$ so that

$$\liminf_{n\to\infty} \inf_{\theta\in\Theta_P} \mathbb{P}\left(\gamma \in CI_\gamma(\alpha_1, \alpha_2)\right) \geq 1 - \alpha_1 - \limsup_{n\to\infty} \sup_{\theta\in\Theta_P} \mathbb{P}(B_\theta).$$

The proof is complete if we can show that $\hat{\sigma}_Y^{-2}\hat{t}^2_{\gamma|\tilde{\Gamma}} \to_w \chi^2_{d-1}$ uniformly over Θ_P since this would imply that $\limsup_{n\to\infty} \sup_{\theta\in\Theta_P} \mathbb{P}(B_\theta) \leq \alpha_2$ by the same argument as in the proof of Theorem 2. Defining $\tilde{\rho}_t = \rho_t - \Sigma_{YX}\Sigma_X^{-1}\tilde{\epsilon}_t$, we have in obvious notation $\hat{t}^2_{\gamma|\tilde{\Gamma}} = S_{\tilde{\rho}\tilde{X}}S^{-1}_{\tilde{X}\tilde{X}}S_{\tilde{X}\tilde{\rho}} + R_n$, where

$$R_n = n\left(r_n S_{\tilde{\epsilon}\tilde{X}} S_{\tilde{X}\tilde{X}} S_{\tilde{X}\tilde{\epsilon}} r_n^T + r_n S_{\tilde{\epsilon}\tilde{X}} S_{\tilde{X}\tilde{X}} S_{\tilde{X}\tilde{\rho}} + S_{\tilde{\rho}\tilde{X}} S_{\tilde{X}\tilde{X}} S_{\tilde{X}\tilde{\epsilon}} r_n^T\right)$$

with $r_n = \hat{\Sigma}_{YX}\hat{\Sigma}_X^{-1} - \Sigma_{YX}\Sigma_X^{-1}$. Since $\hat{\Sigma}$ is uniformly consistent and Σ is uniformly invertible over Θ_P we see that $r_n \to_p 0$ uniformly over Θ_P. This follows from the uniform versions of the continuous mapping theorem and Slutsky's Lemma (see, e.g., Proposition 9 and Proposition 15 in the supplementary material for Lundborg et al. [2022]). Furthermore, all the matrix products in the expression above converge uniformly in distribution over Θ_P by Theorem 1. Thus, $R_n \to_p 0$ uniformly over Θ_P. Now, let $\tilde{\Sigma} \in \mathbb{R}^{d\times d}$ be given by $\tilde{\Sigma}_{11} = \Sigma_Y - \delta^T\Sigma_X\delta$, $(\tilde{\Sigma}_{1i})_{2\leq i\leq d} = (\tilde{\Sigma}_{i1})_{2\leq i\leq d} = 0$, and $(\tilde{\Sigma}_{ij})_{2\leq i,j\leq d} = \Sigma_X$. Similar to (2.12.21), we then find

$$S_{\tilde{\rho}\tilde{X}}S^{-1}_{\tilde{X}\tilde{X}}S_{\tilde{X}\tilde{\rho}} \to_w \left|\left|\left(\tilde{\Sigma}_Y - \tilde{\delta}^T\tilde{\Sigma}_X\tilde{\delta}\right)^{\frac{1}{2}} Z\right|\right|^2$$

uniformly over Θ_P with Z a $(d-1)$-dimensional standard normal random variable. But then, since $\tilde{\Sigma}_Y - \tilde{\delta}^T\tilde{\Sigma}_X\tilde{\delta} = \Sigma_Y - \delta^T\Sigma_X\delta$ which, in turn, is uniformly estimated by $\hat{\sigma}^2_Y$, this completes the proof. □

2.G. Martingale Limit Theorems

We start by stating a strong invariance principle for stationary martingale difference arrays due to Cuny et al. [2021]. A martingale difference array is a doubly infinite array, $(e_{t,n})_{t,n\in\mathbb{N}}$, along with an array of filtrations, $(\mathcal{F}_{t,n})_{t,n\in\mathbb{N}}$, such that, for each n, $e_{t,n}$ is a martingale difference sequence wrt. $\mathcal{F}_{t,n}$.

Theorem 3. *Let $(e_{t,n})_{t,n\in\mathbb{N}}$ be a stationary $\mathbb{R}^d$-valued martingale difference array wrt. $(\mathcal{F}_{t,n})_{t,n\in\mathbb{N}}$ such that $\mathbb{E}(e_{t,n}e_{t,n}^T|\mathcal{F}_{t-1,n}) = \mathbb{E}e_{0,n}e_{0,n}^T = I$ a.s. for all $t \geq 1, n \in \mathbb{N}$ and assume there exists some small $\delta > 0$ such that $\sup_{t,n\in\mathbb{N}} \mathbb{E}||e_{t,n}||^{2+\delta} < \infty$. Then, after possibly enlarging $(\Omega, \mathcal{F}, \mathbb{P})$, there exist a triangular array of random variables, $(\rho_{t,n})_{t\geq 1,n\in\mathbb{N}}$, where each row, $(\rho_{t,n})_{t\in\mathbb{N}}$, is i.i.d. standard normal and such that*

$$\mathbb{E}\left(\left|\left|\sup_{1\leq k\leq n} ||\sum_{t=1}^{k} e_{t,n} - \sum_{t=1}^{k} \rho_{t,n}||\right|\right|\right) = O\left(n^{\frac{1}{2+\delta}} (\log n)^{\frac{1+\delta}{2(2+\delta)}}\right).$$

Proof. The proof is essentially the same as the proof of Theorem 2.1 in Cuny et al. [2021] and we will not go into much detail here, but only mention the key steps and how they generalize to the martingale difference array setting.

First, we note that under above assumptions, Lemma 4.1 in Cuny et al. [2021] holds for triangular arrays as well. Indeed, the constant C depends only on δ, d and $\mathbb{E}||e_{1,n}||^{2+\delta}$ and the latter is uniformly bounded over n.

After possibly enlarging the initial probability space, we can assume that it is large enough to contain a doubly infinite array $(u_{t,n})_{t,n\in\mathbb{N}}$ and a sequence $(\rho_{1,n})_{n\in\mathbb{N}}$ such that, for each fixed n, $u_{t,n}$ is i.i.d. uniform on $[0,1]$ and independent of $e_{t,n}$ and $\rho_{1,n}$ is d-dimensional standard normal independent of $e_{t,n}$ and $u_{t,n}$. Now for each n, we can follow the steps in the proof of Theorem 2.1 in Cuny et al. [2021] and construct a sequence, $(\rho_{t,n})_{t\geq 1}$, of i.i.d. d-dimensional standard normal random variables satisfying certain inequalities. Now define, for $L \in \mathbb{N}$,

$$D_{L,n} = \sup_{l\leq 2^L} \left|\left| \sum_{i=2^L+1}^{2^L+l} e_{n,t} - \rho_{n,t} \right|\right|.$$

By the same arguments as in Cuny et al. [2021], we have, for any $N \in \mathbb{N}$,

$$\sup_{1\leq k\leq 2^{N+1}} \left|\left| \sum_{t=1}^{k} e_{t,n} - \sum_{t=1}^{k} \rho_{t,n} \right|\right| \leq ||e_{1,n} - \rho_{1,n}|| + \sum_{L=0}^{N-1} D_{L,n} + D_{N,n}.$$

Furthermore, they show that the array, $\rho_{t,n}$, was constructed in such a way that there exists a constant c_0 depending only on δ, d and $\mathbb{E}||e_{1,n}||^{2+\delta}$ such that

$$||D_{L,n}||_1 \leq C 2^{\frac{L}{2+\delta}} L^{\frac{1+\delta}{2(2+\delta)}}$$

for all $L \in \mathbb{N}$. Since $\mathbb{E}||e_{1,n}||^{2+\delta}$ is uniformly bounded in n, we may assume that c_0

depends only on δ and d. For $2^N \leq n < 2^{N+1}$, we have

$$\begin{aligned}
\sum_{L=1}^{N-1} 2^{\frac{L}{2+\delta}} L^{\frac{1+\delta}{2(2+\delta)}} &\leq \left(\frac{\log n}{\log 2}\right)^{\frac{1+\delta}{2(2+\delta)}} \sum_{L=0}^{N-1} 2^{\frac{L}{2+\delta}} \\
&= \left(\frac{\log n}{\log 2}\right)^{\frac{1+\delta}{2(2+\delta)}} \frac{1-2^{\frac{N}{2+\delta}}}{1-2^{\frac{1}{2+\delta}}} \\
&\leq \frac{1}{\left(1-2^{\frac{1}{2+\delta}}\right)(\log 2)^{\frac{1+\delta}{2(2+\delta)}}} (\log n)^{\frac{1+\delta}{2(2+\delta)}} \left(n^{\frac{1}{2+\delta}}+1\right) \\
&\leq c_1 n^{\frac{1}{2+\delta}} (\log n)^{\frac{1+\delta}{2(2+\delta)}}
\end{aligned}$$

where c_1 does not depend on n. Finally, under the assumptions of the theorem, there exists a constant c_2 not depending on n and such that

$$\mathbb{E}\,||e_{1,n} - \rho_{1,n}|| \leq c_2.$$

Putting all the pieces together, we find that

$$\mathbb{E}\left(\left|\left|\sup_{1\leq k\leq n} ||\sum_{t=1}^{k} e_{t,n} - \sum_{t=1}^{k} \rho_{t,n}||\right|\right|\right) \leq c_2 + c_0(c_1+1) n^{\frac{1}{2+\delta}} (\log n)^{\frac{1+\delta}{2(2+\delta)}}$$

which is the result we wanted. □

Next we adapt the weak law of large numbers from Theorem 6 in De Jong [1998] to the multidimensional setting.

Theorem 4. *Let $(e_{t,n})_{t,n\in\mathbb{N}}$ be an $\mathbb{R}^d$-valued martingale difference array wrt. $(\mathcal{F}_{t,n})_{t,n\in\mathbb{N}}$. Assume there exists $\delta > 0$ such that $\sup_{t,n} \mathbb{E}||e_{t,n}||^{1+\delta} < \infty$. Then, for any $\epsilon > 0$, it holds that*

$$\left|\left|\frac{1}{n}\sum_{t=1}^{n} e_{t,n}\right|\right| = o_p\left(n^{\frac{1}{1+\delta}-1+\epsilon}\right).$$

Proof. Fix $\epsilon > 0$ and let $a \in \mathbb{R}$ and $\tilde{e}_{t,n} = a^T e_{t,n}$. By Cauchy-Schwartz, we have

$$\sup_{t,n\in\mathbb{N}} |\tilde{e}_{t,n}|^{1+\delta} \leq ||a||^{1+\delta} \sup_{t,n\in\mathbb{N}} ||e_{t,n}||^{1+\delta} = C < \infty.$$

Now, define $k_n = n^{\frac{1}{1+\delta}+\epsilon}$. Then, with $p = 1+\delta$,

$$k_n^{-p} \sum_{t=1}^{n} (\mathbb{E}|\tilde{e}_{t,n}|)^p \leq C k_n^{-p} n = o(1)$$

and, by Theorem 6 in De Jong [1998],

$$\left|\frac{1}{n}\sum_{t=1}^{n} \tilde{e}_{t,n}\right| = o_p\left(n^{\frac{1}{1+\delta}-1+\epsilon}\right).$$

The result then follows since a was arbitrary. □

Finally, we need the following well-known martingale difference array central limit theorem (see, e.g., Theorem 1 of Chapter VIII in Pollard [1984]).

Theorem 5. *Let $(e_{t,n})_{t,n\in\mathbb{N}}$ be an $\mathbb{R}^d$-valued martingale difference array wrt. $(\mathcal{F}_{t,n})_{t,n\in\mathbb{N}}$. Assume that*

$$\sum_{t=1}^{n} \mathbb{E}\left(e_{t,n}e_{t,n}^T|\mathcal{F}_{t-1,n}\right) \to_p I$$

and, for each $\gamma > 0$,

$$\sum_{t=1}^{n} \mathbb{E}\left(||e_{t,n}||^2 \mathbf{1}\left(||e_{t,n}|| > \gamma\right)|\mathcal{F}_{t-1,n}\right) \to_p 0$$

for $n \to \infty$. Then, $\sum_{t=1}^{n} e_{t,n} \to_w \mathcal{N}(0, I)$ for $n \to \infty$.

2.H. Gaussian Approximation

In this section we detail how the Gaussian approximation described in Section 3.1 is achieved. Throughout we assume that $X_{t,\theta}$ and $\epsilon_{t,\theta}$ satisfy Assumptions U and M with $F_\theta = I$. The key result is the strong invariance principle of Theorem 3. Although it is stated in terms of martingale difference arrays, the version we need (Lemma 1 below) follows easily from Proposition 8 in Lundborg et al. [2022] and Assumptions M.3, U.1, and U.2.

Lemma 1. *We can enlarge the initial probability space such that there exists a family of stochastic processes $(\rho_{t,\theta})_{t\geq 1,\theta\in\Theta}$ where, for each θ, the sequence $\rho_{t,\theta}$ is i.i.d. d-dimensional gaussian with mean 0 and covariance matrix Σ and such that*

$$\sup_{\theta\in\Theta}\sup_{1\leq k\leq n}\left|\left|\sum_{t=1}^{k}\epsilon_{t,\theta} - \sum_{t=1}^{k}\rho_{t,\theta}\right|\right| = o_p\left(n^{\frac{1}{2}-\beta}\right).$$

for some $\beta > 0$.

Lemma 2. *For any $\epsilon > 0$, we have*

$$n^{-\frac{1}{1+\delta}-\epsilon}\sum_{t=1}^{n}\left(\epsilon_{t,\theta}\epsilon_{t,\theta}^T - \Sigma\right) \to_p 0$$

uniformly over Θ.

Proof. Fix $a, b \in \mathbb{R}^d$ and define $\xi_{t,\theta} = a^T(\epsilon_{t,\theta}\epsilon_{t,\theta}^T - \Sigma)b$ so that $\xi_{t,\theta}$ is a one-dimensional martingale difference sequence for each $\theta \in \Theta$. By Cauchy-Schwartz, we have, for all $t \in \mathbb{N}$ and $\theta \in \Theta$, $|\xi_{t,\theta}| \leq ||a\Sigma^{\frac{1}{2}}||||b\Sigma^{\frac{1}{2}}||\left(||\Sigma^{-\frac{1}{2}}\epsilon_{t,\theta}||^2 + 1\right)$ so that, by assumption,

$$\sup_{\theta\in\Theta}\mathbb{E}|\xi_{t,\theta}|^{1+\frac{\delta}{2}} < \infty.$$

The result then follows from Theorem 4 in combination with Proposition 8 in Lundborg et al. [2022]. □

The following is similar to Lemma 4 in Mikusheva [2007]. It shows that many of the important statistics can be replaced by their Gaussian counterpart.

Lemma 3. *There exists $\beta > 0$ such that*

(a) $\sup_{\theta\in\Theta} \sup_{1\le t\le n} ||X_{t,\theta}/\sqrt{n} - Y_{t,\theta}/\sqrt{n}|| = o_p(n^{-\beta})$,

(b) $\sup_{\theta\in\Theta} \sup_{1\le t\le n} \{||X_{t,\theta}/\sqrt{n}|| + ||Y_{t,\theta}/\sqrt{n}||\} = O_p(1)$,

(c) $\sup_{\theta\in\Theta} ||\sum_{t=1}^{n}\sum_{s=1}^{t} \epsilon_{s,\theta}\epsilon_{t,\theta}^T/n - \rho_{s,\theta}\rho_{t,\theta}^T/n|| = o_p(n^{-\beta})$.

(d) $\sup_{\theta\in R_{n,d}} ||H^{-\frac{1}{2}}(S_{XX} - S_{YY})H^{-\frac{1}{2}}|| = o_p(n^{1-\eta-\beta})$.

(e) $\sup_{\theta\in R_{n,d}} ||\sqrt{n}H^{-\frac{1}{2}}(S_{X\epsilon} - S_{Y\rho})|| = o_p(n^{\frac{3}{2}(1-\eta)-\beta})$.

Proof. For the proof of part *(a)* we use summation by parts to write

$$\begin{aligned} X_{t,\theta} &= \sum_{s=1}^{t} \Gamma^{t-s}\epsilon_{s,\theta} = \sum_{s=1}^{t} \epsilon_{s,\theta} - \sum_{s=1}^{t-1} \left(\Gamma^{t-s+1} - \Gamma^{t-s}\right) \sum_{k=1}^{s+1} \epsilon_{k,\theta} \\ &= \sum_{s=1}^{t} \epsilon_{s,\theta} - (\Gamma - I) \sum_{s=1}^{t-1} \Gamma^{t-s} \sum_{k=1}^{s+1} \epsilon_{k,\theta} \end{aligned}$$

and a similar expression holds for $Y_{t,\theta}$. Thus,

$$\begin{aligned} \sup_{\theta\in\Theta} \sup_{1\le t\le n} \frac{1}{\sqrt{n}} ||X_{t,\theta} - Y_{t,\theta}|| \le \sup_{\theta\in\Theta} \sup_{1\le t\le n} \left|\left|(\Gamma - I) \sum_{s=1}^{t-1} \Gamma^{t-s} + I\right|\right| \\ \times \frac{1}{\sqrt{n}} \sup_{\theta\in\Theta} \sup_{1\le t\le n} \left|\left|\sum_{s=1}^{t} \epsilon_{s,\theta} - \sum_{s=1}^{t} \rho_{s,\theta}\right|\right|. \end{aligned}$$

Since the first term on the right hand side is bounded by Assumptions U.3 and U.4, Lemma 1 yields *(a)*.

To prove *(b)*, we start with the same expression for $Y_{t,\theta}$ as above. This gives us

$$\sup_{\theta\in\Theta} \sup_{1\le t\le n} \frac{Y_{t,\theta}}{\sqrt{n}} \le \sup_{\theta\in\Theta} \sup_{1\le t\le n} \left|\left|(\Gamma - I) \sum_{s=1}^{t-1} \Gamma^{t-s} + I\right|\right| \frac{1}{\sqrt{n}} \sup_{\theta\in\Theta} \sup_{1\le t\le n} \left|\left|\sum_{s=1}^{t} \rho_{s,\theta}\right|\right|.$$

Again, the first term on the right hand side is bounded uniformly over n. For the second term, we have

$$\begin{aligned} \frac{1}{\sqrt{n}} \sup_{\theta\in\Theta} \sup_{1\le t\le n} \left|\left|\sum_{s=1}^{t} \rho_{s,\theta}\right|\right| &\le \sup_{\theta\in\Theta} \left|\left|\Sigma^{\frac{1}{2}}\right|\right| \sup_{1\le t\le n} \left|\left|\frac{1}{\sqrt{n}} \sum_{s=1}^{t} \Sigma^{-\frac{1}{2}}\rho_{s,\theta}\right|\right| \\ &\le C \sup_{\theta\in\Theta} \sup_{1\le t\le n} \left|\left|\frac{1}{\sqrt{n}} \sum_{s=1}^{t} \Sigma^{-\frac{1}{2}}\rho_{s,\theta}\right|\right| = O_p(1) \end{aligned}$$

since $\Sigma^{-\frac{1}{2}}\rho_{s,\theta}$ is i.i.d. standard normal for all $\theta \in \Theta$. The result then follows from *(a)*.

For *(c)* we start with

$$\frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{t}\epsilon_{s,\theta}\epsilon_{t,\theta}^{T} = \frac{1}{2n}\left(\sum_{t=1}^{n}\epsilon_{t,\theta}\right)\left(\sum_{t=1}^{n}\epsilon_{t,\theta}\right)^{T} + \frac{1}{2n}\sum_{t=1}^{n}\left(\epsilon_{t,\theta}\epsilon_{t,\theta}^{T} - \Sigma\right) + \frac{1}{2}\Sigma.$$

Lemma 2 then yields

$$\sup_{\theta\in\Theta}\left|\left|\frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{t}\epsilon_{s,\theta}\epsilon_{t,\theta}^{T} - \frac{1}{2n}\left(\sum_{t=1}^{n}\epsilon_{t,\theta}\right)\left(\sum_{t=1}^{n}\epsilon_{t,\theta}\right)^{T} - \frac{1}{2}\Sigma\right|\right| = o_p(n^{-\beta})$$

and a similar argument holds for $\frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{t}\rho_{s,\theta}\rho_{t,\theta}^{T}$. Thus,

$$\begin{aligned}\sup_{\theta\in\Theta}\left|\left|\frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{t}\epsilon_{s,\theta}\epsilon_{t,\theta}^{T} - \rho_{s,\theta}\rho_{t,\theta}^{T}\right|\right| &= \sup_{\theta\in\Theta}\left|\left|\frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{n}\epsilon_{s,\theta}\epsilon_{t,\theta}^{T} - \rho_{s,\theta}\rho_{t,\theta}^{T}\right|\right| + o_p(n^{-\beta}) \\ &\leq C_n \sup_{\theta\in\Theta}\left|\left|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\epsilon_{t,\theta} - \rho_{t,\theta}\right|\right| + o_p(n^{-\beta})\end{aligned}$$

where $C_n = \sup_{\theta\in\Theta}\left\{\left|\left|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\epsilon_{t,\theta}\right|\right| + \left|\left|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\rho_{t,\theta}\right|\right|\right\}$. Since the law of $\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\rho_{t,\theta}$ is equal to a d-dimensional Gaussian with mean 0 and covariance matrix Σ, Assumption U.2 yields $\sup_{\theta\in\Theta}\left|\left|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\rho_{t,\theta}\right|\right| = O_p(1)$ and, by Lemma 1, we get $C_n = O_p(1)$ and therefore also the result in *(c)*.

To prove *(d)*, we first note that

$$\sup_{\theta\in\Theta}\frac{1}{n}\left|\left|S_{XX} - S_{YY}\right|\right| \leq C_n \sup_{\theta\in\Theta}\sup_{1\leq t\leq n}\frac{1}{\sqrt{n}}\left|\left|X_{t,\theta} - Y_{t,\theta}\right|\right|$$

where $C_n = \sup_{\theta\in\Theta}\frac{1}{\sqrt{n}}\left(\sup_{1\leq t\leq n}\left|\left|X_{t,\theta}\right|\right| + \sup_{1\leq t\leq n}\left|\left|Y_{t,\theta}\right|\right|\right)$. From Lemma 1, we find that

$$\sup_{\theta\in R_{n,d}}\sigma_{max}\left(H^{-\frac{1}{2}}\right) = \sup_{\theta\in R_{n,d}}\left(\sigma_{min}(H)\right)^{-\frac{1}{2}} = O(n^{-\frac{\eta}{2}}) \tag{2.H.1}$$

so, by part *(a)* and *(b)*, we get the result in *(d)*.

For the proof of *(e)* we again use summation by parts to write

$$\begin{aligned}S_{X\epsilon} &= \frac{1}{n}\left(X_{n-1,\theta}\sum_{t=1}^{n}\epsilon_{t,\theta}^{T} - \sum_{t=1}^{n-1}\left(X_{t,\theta} - X_{t-1,\theta}\right)\sum_{s=1}^{t}\epsilon_{s,\theta}^{T}\right) \\ &= \frac{1}{n}\left(X_{n-1,\theta}\sum_{t=1}^{n}\epsilon_{t,\theta}^{T} - (\Gamma - I)\sum_{t=1}^{n-1}\sum_{s=1}^{t}X_{t-1,\theta}\epsilon_{s,\theta}^{T} - \sum_{t=1}^{n-1}\sum_{s=1}^{t}\epsilon_{t,\theta}\epsilon_{s,\theta}^{T}\right),\end{aligned}$$

and similarly for $S_{Y\rho}$. We have, for all $1 \leq t \leq n$,

$$\sup_{\theta\in\Theta}\left|\left|\frac{1}{n}\sum_{s=1}^{t} X_{t-1,\theta}\epsilon_{s,\theta}^T - Y_{t-1,\theta}\rho_{s,\theta}^T\right|\right| \leq C_n \sup_{\theta\in\Theta}\left\{\sup_{1\leq k\leq n}\frac{1}{\sqrt{n}}\left(||X_{k,\theta}-Y_{k,\theta}|| + \left|\left|\sum_{s=1}^{k}\epsilon_{s,\theta}-\rho_{s,\theta}\right|\right|\right)\right\},$$

where $C_n = \sup_{\theta\in\Theta}\left\{\sup_{1\leq k\leq n}||\frac{1}{\sqrt{n}}X_{k,\theta}|| + \sup_{1\leq k\leq n}\left|\left|\frac{1}{\sqrt{n}}\sum_{s=1}^{k}\rho_{s,\theta}\right|\right|\right\} = O_p(1)$ by part *(b)*. From part *(a)* and Lemma 1 it then follows that

$$\sup_{\theta\in\Theta}\left|\left|\frac{1}{n}\sum_{s=1}^{t} X_{t-1,\theta}\epsilon_{s,\theta}^T - Y_{t-1,\theta}\rho_{s,\theta}^T\right|\right| = o_p\left(n^{-\beta}\right).$$

We also have $\sup_{\theta\in R_{n,d}} n||(\Gamma - I)|| = o_p(n^{1-\eta})$ so that

$$\sup_{\theta\in R_{n,d}}\left|\left|\frac{1}{n}(\Gamma - I)\sum_{t=1}^{n-1}\sum_{s=1}^{t} X_{t-1,\theta}\epsilon_{s,\theta}^T - Y_{t-1,\theta}\rho_{s,\theta}\right|\right| = o_p\left(n^{1-\eta-\beta}\right)$$

and, by part *(c)*, $\sup_{\theta\in R_{n,d}}||S_{X\epsilon} - S_{Y\rho}|| = o_p\left(n^{1-\eta-\beta}\right)$. The result then follows from (2.H.1). □

Lemma 1 allows us to assume that $\epsilon_{t,\theta}$ is an i.i.d. Gaussian sequence with mean zero and covariance matrix Σ for each $\theta \in R_{n,d}$. Indeed, let $\rho_{t,\theta}$ be as given in the Lemma and define the family $(Y_{t,\theta})_{t\in\mathbb{N},\theta\in\Theta}$ by

$$Y_{t,\theta} = \Gamma Y_{t-1,\theta} + \rho_{t,\theta}, \quad Y_{0,\theta} = 0$$

and define the corresponding sample covariances

$$S_{YY} = \frac{1}{n}\sum_{t=1}^{n} Y_{t-1,\theta}Y_{t-1,\theta}^T, \quad S_{Y\rho} = \frac{1}{n}\sum_{t=1}^{n} Y_{t-1,\theta}\rho_{t,\theta}^T.$$

Then, by Lemma 3, we can pick η close enough to 1 such that

$$\sup_{\theta\in R_{n,d}}\left\{\left|\left|H^{-\frac{1}{2}}(S_{XX} - S_{YY})H^{-\frac{1}{2}}\right|\right| + \left|\left|\sqrt{n}H^{-\frac{1}{2}}(S_{X\epsilon} - S_{Y\rho})\right|\right|\right\} = o_p(1). \tag{2.H.2}$$

2.I. Simulations

2.I.1 Confidence intervals

For each $n \in \{50, 75, 100\}$ and $d \in \{3, 4, 5\}$ the simulation experiment is repeated 1000 times. In each repetition, for $i, j = 1, ..., d$, we draw $U_{ij} \sim \text{Unif}([0, 1])$ and set $\Gamma =$

$U^{-1}\Lambda_n U$ where $\Lambda_n \in \mathbb{R}^{d\times d}$ is diagonal with $\Lambda_{n,11} = 1$ and $\Lambda_{n,ii} = 1 - (1/n)^{1/(i-1)}$ for $i = 2, ..., d$. We then sample $\epsilon_t \sim \mathcal{N}(0, \Sigma)$ i.i.d. for $t = 1, ..., n$ with

$$\Sigma = \frac{1}{2}\left(I + \mathbb{1}\mathbb{1}^T\right)$$

where $\mathbb{1} = (1, ..., 1)^T \in \mathbb{R}^d$ and let

$$X_t = \Gamma X_{t-1} + \epsilon_t \text{ for } t = 1, ..., n, \quad X_0 = 0.$$

X_t is a sample from a VAR(1) process with $\theta = (\Gamma, \Sigma, \cdot)$. We then compute CI_b, CI_{IV}, and CI_{LA} for this sample and record the length of each confidence interval and whether it contains Γ_{11}.

2.I.2 Predictive regression testing

For both simulation experiments we fix $d = 4$ and $\alpha = 0.1$. This implies that $\tilde{\Gamma} \in \mathbb{R}^{3\times 3}$. The two regimes correspond to two different choices of $\tilde{\Gamma}$:

- *Mixed Regime*: In this setting $\tilde{\Gamma}$ is chosen as above, that is, with roots of differing proximity to unity and with random eigenvectors sampled anew for every simulation run.
- *Non-stationary Regime*: In this setting we set $\tilde{\Gamma} = I$ so that $\tilde{X}_t$ is a random walk.

In both regimes the errors are i.i.d. Gaussian with covariance matrix Σ as given above.

To obtain Figure 1, we do the following: For each $n \in \{10, 20, ..., 200\}$, we draw two samples X_t from the VAR(1) processes given by the two choices of $\tilde{\Gamma}$ and under the null $H_0 : \gamma = 0$. We then compute the three tests on both samples recording whether the null was rejected or not. This is repeated 1000 times and the rejection rate is the proportion of times the null was rejected across all simulations.

For Figure 2, we do essentially the same thing except that we now fix $n = 100$ and perform the experiment across different choices of $\gamma \neq 0$. In particular, we run the experiment for $\gamma = \delta\mathbb{1}$, $\delta \in \{0.005, 0.01, ..., 0.1\}$ and record the proportion of times the null was rejected across all 1000 simulations.

2.I.3 EAM

We present a slightly generalized version of the EAM-algorithm. EAM stands for Evaluation-Approximation-Maximization and the algorithm can more or less be split into three steps. It is an algorithm for solving problems of the following form

$$\begin{aligned} &\sup f(x) \\ &\text{s.t. } g(x) \leq c(x) \end{aligned}$$

over $x \in \mathcal{X} \subset \mathbb{R}^p$ where f, g, and c are fixed scalar functions sufficiently smooth and satisfying certain requirements. In Kaido et al. [2019] it is required that $f(x) = v^T x$

for some $v \in \mathbb{R}^p$. Here we only require that $f(x)$ is convex and twice continuously differentiable on $\mathcal{X}$. We assume that c is costly to evaluate. Without going into too much detail, the algorithm proceeds as follows:

1. *Initialization:* Randomly sample initial points $x^{(1)}, ..., x^{(k)}$ from $\mathcal{X}$ and evaluate $c(x^{(i)})$ for $i = 1, ..., k$. Set $L = k$.

2. Iterate the following three steps until convergence:
   a) *E-step*: Evaluate $c(x^{(L)})$ and pick current optimum
   $$y^{*,L} = \max\left\{f\left(x^{(i)}\right) : g\left(x^{(i)}\right) \leq c\left(x^{(i)}\right), i = 1, ..., L\right\}$$
   b) *A-step*: Approximate $x \mapsto c(x)$ by a Gaussian process regression model, with mean μ, constant variance σ^2, and covariance kernel $K_\beta(x - x') = \exp(-\sum_{i=1}^p |x_i - x_i'|^2/\beta_i)$, fitted on $(c(x^{(i)}), x^{(i)})$, $i = 1, ..., L$. Fitting the model yields the mean function $c_L(x)$ and the variance function $s_L(x)$ as well as the fitted parameters $\hat{\mu}_L, \hat{\sigma}_L, \hat{\beta}_L$.
   c) *M-step*: With probability $1 - \epsilon$, let
   $$x^{(L+1)} = \arg\max_{x \in \mathcal{X}} EI_L(x)$$
   and with probability ϵ draw $x^{(L+1)}$ randomly from $\mathcal{X}$. Set $L = L + 1$.

$EI_L(x)$ is the *expected improvement function* and it is given by

$$EI_L(x) = \left(f(x) - y^{*,L}\right)_+ \left(1 - \Phi\left(\frac{g(x) - c_L(x)}{\hat{\sigma}_L s_L(x)}\right)\right)$$

where $\cdot_+ = \max\{\cdot, 0\}$ and Φ is the standard normal CDF. Note that the optimization problem in the M-step can be reformulated as a constrained optimization problem with smooth objective function and smooth constraints for which all derivatives are known and it can therefore be solved with standard solvers. A key observation is that we only evaluate c once per iteration. In practice this results in far fewer evaluations of c when compared to, say, grid methods.

This algorithm was intended to compute confidence intervals that arise as projections of confidence regions exactly as is the case for CI_b. In this case we would simply take $x = \Gamma$ and let $f(\Gamma) = \Gamma_{11}$, $g(\Gamma) = \hat{t}^2_\Gamma$, and $c(\Gamma) = \tilde{q}_{n,\Gamma}(\alpha)$. It does not really matter that Γ is a matrix since we can just vectorize it and redefine all the functions correspondingly. This would give us the upper bound of CI_b and the lower bound can be found by taking $f(\Gamma) = -\Gamma_{11}$.

Similarly, the EAM-algorithm can be used to compute φ_b. This is done by letting $x = \tilde{\Gamma}$ and then $f(\tilde{\Gamma}) = \hat{t}^2_{0|\tilde{\Gamma}}$ with g and c as before, but for $\tilde{\Gamma}$ instead of Γ. Note that in both cases f and g are polynomials of $\text{vec}(x)$ and are therefore smooth with known derivatives of all orders.[11]

[11] All code used for the simulations can be found at https://github.com/cholberg/unif_inf_var. Our implementation of the EAM-algorithm is based on Kaido et al. [2017].

2.J. Lag augmentation

We shall prove that the lag augmented estimator converges in distribution to a normal distribution uniformly over Θ upon which it follows that standard inference is uniformly valid by the same arguments as applied in the proof of Theorem 2.

Lemma 1. *Γ_{LA} be defined as in Section 2.12.2. Assume U and M. Then, as $n \to \infty$,*

$$\sqrt{n}\text{vec}\left(\hat{\Gamma}_{LA} - \Gamma\right) \to_w \mathcal{N}\left(0, \Sigma^{-1} \otimes \Sigma\right)$$

uniformly over Θ.

Proof. We write $S_{\epsilon X'} = \sum_{t=2}^{n} \epsilon_t X_{t-2}^T/n$, $S_{X'X'} = \sum_{t=2} X_{t-2}X_{t-2}^T/n$, $S_{\epsilon'X'} = \sum_{t=2} \epsilon_{t-1}X_{t-2}^T/n$, $S_{\epsilon'\epsilon'} = \sum_{t=2}^{n} \epsilon_{t-1}\epsilon_{t-1}^T$, and $S_{\epsilon\epsilon'} = \sum_{t=2}^{n} \epsilon_t\epsilon_{t-1}^T$. Then,

$$\begin{aligned}
\hat{\Gamma}_{LA} - \Gamma &= S_{\epsilon\bar{X}}S_{\bar{X}\bar{X}}^{-1}D \\
&= \left(S_{\epsilon X} - S_{\epsilon X'}S_{X'X'}^{-1}\left(S_{X'X'}\Gamma^T + S_{X'\epsilon'}\right)\right)\left(S_{\epsilon'\epsilon'} - S_{\epsilon'X'}S_{X'X'}^{-1}S_{X'\epsilon'}\right)^{-1} \\
&= \left(S_{\epsilon\epsilon'} - S_{\epsilon X'}S_{X'X'}^{-1}S_{X'\epsilon'}\right)\left(S_{\epsilon'\epsilon'} - S_{\epsilon'X'}S_{X'X'}^{-1}S_{X'\epsilon'}\right)^{-1}
\end{aligned}$$

where we used the relation $X_{t-1} = \Gamma X_{t-2} + \epsilon_{t-1}$ multiple times. By Theorem 1, $S_{\epsilon X'}S_{X'X'}^{-1}S_{X'\epsilon'} \to_p 0$ and $S_{\epsilon'X'}S_{X'X'}^{-1}S_{X'\epsilon'} \to_p 0$ for $n \to \infty$ uniformly over Θ. Furthermore, Theorem 4 and 5 in the Appendix yield $S_{\epsilon'\epsilon'} \to_p \Sigma$ and $\sqrt{n}\text{vec}(S_{\epsilon\epsilon'}) \to_w \mathcal{N}(0, \Sigma \otimes \Sigma)$ for $n \to \infty$ uniformly over Θ. Finally, since Σ is uniformly invertible and bounded on Θ, the uniform versions of the continuous mapping theorem and Slutsky's Lemma (Proposition 9 and Proposition 15 in Lundborg et al. [2022]), yield the desired result. □

2.K. IVX

We shall prove that the IVX t^2-statistic converges in distribtuion to $\chi^2_{d^2}$ uniformly over Θ after which the rest follows by the same arguments as those applied in the proof of Theorem 2.

Theorem 6. *Assume that Assumptions M and U are true. Let $\hat{\Gamma}_{IV}$ be the IVX estimator and $\hat{t}^2_{IV,\Gamma}$ the corresponding t^2-statistic as defined in Section 2.12.3. Then, for $n \to \infty$,*

$$\hat{t}^2_{IV,\Gamma} \to \chi^2_{d^2}$$

uniformly over Θ.

The proof of Theorem 6 that we present here is conceptually different from the proofs presented so far. We rely on the theory developed in Magdalinos and Phillips [2020], Phillips et al. [2009], but since we do not require that all roots approach unity at the same rate, there are some extra difficulties that need to be dealt with. In particular, we need to employ a different normalization in obtaining the asymptotics of S_{ZZ} and $S_{\epsilon Z}$. Furthermore, Theorem 6 shows that the suggested IVX approach is truly uniformly valid (at least over the suggested parameter space, Θ). The first lemma is of a technical nature.

Lemma 1. *Let $(\theta_n)_{n\in\mathbb{N}} \subset \Theta$ satisfy Assumptions M and U with $F_{\theta_n} = I$ and $\beta \in (0,1)$. Then, there exist $0 \leq r \leq d$, $(k_n)_{n\in\mathbb{N}} \subset \mathbb{N}$ strictly increasing, and $(\tilde{\theta})_{n\in\mathbb{N}} \subset \Theta$ such that*

(i) $\theta_{k_n} = \tilde{\theta}_{k_n}$, $\forall n \in \mathbb{N}$,

(ii) $n^{\beta}(1 - \tilde{\Gamma}_{n,ii}) \to \kappa_i \in \mathbb{C}$, $|\kappa_i| \in [0,1]$, *for* $1 \leq i \leq r$,

(iii) $n^{-\beta}(1 - \tilde{\Gamma}_{n,ii})^{-1} \to \kappa_i \in \mathbb{C}$, $|\kappa_i| \in [0,1]$, *for* $r+1 \leq i \leq d$,

(iv) $\tilde{\theta}_n \to \theta \in \Theta$,

where $\tilde{\theta}_n = (\tilde{\Gamma}_n, \tilde{\Sigma}_n, \cdot)$ and all limits are taken as $n \to \infty$.

Proof. Fix $(\theta_n)_{n\in\mathbb{N}} \subset \Theta$ and $\beta \in (0,1)$. For each $n \in \mathbb{N}$, let $0 \leq r_n \leq d$ be such that $|n^{\beta}(1 - \Gamma_{n,ii})| \leq 1$ for $1 \leq i \leq r_n$ and $|n^{-\beta}(1 - \Gamma_{n,ii})^{-1}| \leq 1$ otherwise. Then, $(r_n)_{n\in\mathbb{N}}$ is a sequence in $\{0, 1, ..., d\}$ so that, by compactness, it has a convergent sub-sequence. In other words, there exists a sub-sequence, $(\theta_{n_k})_{k\in\mathbb{N}}$, and $0 \leq r \leq d$ such that $|(n_k)^{\beta}(1 - \Gamma_{n_k,ii})| \leq 1$ for $1 \leq i \leq r$ and $|(n_k)^{-\beta}(1 - \Gamma_{n_k,ii})^{-1}| \leq 1$ otherwise. By Bolzano-Weierstrass, we may assume without loss of generality (passing to another sub-sequence if necessary) that there exists $\kappa \in \mathbb{C}^d$ with $|\kappa_i| \leq 1$ and such that

$$(n_k)^{\beta}(1 - \Gamma_{n_k,ii}) \to \kappa_i, \text{ for } 1 \leq i \leq r$$
$$(n_k)^{-\beta}(1 - \Gamma_{n_k,ii})^{-1} \to \kappa_i, \text{ for } r+1 \leq i \leq d$$

for $k \to \infty$. By another compactness argument, we can furthermore choose the sub-sequence such that $\theta_{n_k} \to \theta = (\Gamma, \Sigma, c) \in \Theta$. Now, take some $\delta \in (0, \beta)$, let $0 \leq r_1 \leq r_2 \leq d$ be such that $|\kappa_i| > 0$ for $r_1 < i \leq r_2$ and $\kappa_i = 0$ otherwise, and define the diagonal matrix $C_n \in \mathbb{C}^{d\times d}$ by

$$C_{n,ii} = \begin{cases} n^{-\delta}, & \text{if } i \leq r_1, \\ \kappa_i, & \text{if } r_1 < i \leq r, \\ \kappa_i^{-1}, & \text{if } r < i \leq r_2, \\ n^{\delta}, & \text{otherwise.} \end{cases}$$

By Assumptions U.3 and U.4, we must have $(\Gamma_{i,j})_{1\leq i,j\leq r_2} = I_{r_2}$ so we find that $\Gamma'_n = \Gamma - n^{-\beta}C_n$ satisfies (ii) and (iii) with $\Gamma'_n \to \Gamma$ for $n \to \infty$. Finally, let $\tilde{\Gamma}_n = \Gamma_{n_k}$ if $n = n_k$ for some $k \in \mathbb{N}$ and $\tilde{\Gamma}_n = \Gamma'_n$ otherwise and $\tilde{\Sigma}_n = \Sigma_{n_k}$ and $\tilde{c}_n = c_{n_k}$ for $n_k \leq n < n_{k+1}$. Then, $\tilde{\theta}_n = (\tilde{\Gamma}_n, \tilde{\Sigma}_n, \tilde{c}_n)$ satisfies all the conditions. □

Sequences of parameters like $\tilde{\theta}_n$ in the above Lemma fit nicely into the framework of Magdalinos and Phillips [2020], Phillips et al. [2009]. We can adapt their results to this more general setup. Fix some $\beta \in (0,1)$ and consider a sequence $(\theta_n)_{n\in\mathbb{N}} \subset \Theta$ such that conditions (ii) and (iii) are satisfied for some $0 \leq r \leq d$ and $\kappa \in \mathbb{C}^d$ with $|\kappa_i| \leq 1$. For such a sequence, we can define the integers $0 \leq r_1 \leq r_2 \leq d$ as in the proof above along with the diagonal matrices $D_n \in \mathbb{C}^{d\times d}$ given by $D_{n,ii} = n^{-\beta}$ for $1 \leq i \leq r_2$ and $D_{n,ii} = (1 - |\Gamma_{n,ii}|)$ otherwise. This normalization is sufficiently flexible to ensure convergence of the relevant sample covariance matrices.

Lemma 2. *Let $\beta \in (\frac{1}{2}, 1)$ and $(\theta_n)_{n\in\mathbb{N}} \subset \Theta$ be a sequence of parameters satisfying Assumptions M and U with $F_{\theta_n} = I$ as well as (ii), (iii), and (iv) of Lemma 1 for some $0 \leq r \leq d$, $\kappa \in \mathbb{C}^d$ with $|\kappa_i| \leq 1$, and $\theta \in \Theta$. For $(D_n)_{n\in\mathbb{N}}$ as defined above and $\mathrm{vec}(V) \sim \mathcal{N}(0, I)$, there exists a sequence of positive definite matrices $(\Sigma_{Z,n})_{n\in\mathbb{N}}$ such that*

$$\limsup_{n\to\infty} \left\{ \sigma_{min}\left(\Sigma_{Z,n}\right)^{-1} + \sigma_{max}\left(\Sigma_{Z,n}\right) \right\} < \infty$$

and the following holds for any $\epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\left| D_n^{\frac{1}{2}} S_{ZZ} D_n^{\frac{1}{2}} - \Sigma_{Z,n} \right|\right| > \epsilon \right) = 0, \tag{2.K.1}$$

$$\lim_{n\to\infty} d_{BL}\left(\sqrt{n}\Sigma_{Z,n}^{-\frac{1}{2}} D_n^{\frac{1}{2}} S_{Z\epsilon} \Sigma^{-\frac{1}{2}}, V \right) = 0. \tag{2.K.2}$$

The following result is useful for the proof of Lemma 2. With a slight abuse of notation, for any $\theta \in \Theta$, let $(I-\Gamma)^{\frac{1}{2}}$ be the diagonal matrix given by the principal square root of the diagonal of $I - \Gamma$.

Lemma 3. *Assume that Assumptions M and U hold with $F_\theta = I$. Then,*

$$\sup_{\theta\in\Theta} \sup_{t\geq 1} \left|\left| \mathbb{E}\left((I-\Gamma)^{\frac{1}{2}} X_{t,\theta} X_{t,\theta}^T (I-\Gamma)^{\frac{1}{2}} \right) \right|\right| < \infty \tag{2.K.3}$$

and, furthermore,

$$\lim_{n\to\infty} \sup_{\theta\in R_{n,0}} \left|\left| (I-\Gamma)^{\frac{1}{2}} H (I-\Gamma)^{\frac{1}{2}} - \Sigma_X \right|\right| = 0 \tag{2.K.4}$$

where $\mathrm{vec}(\Sigma_X) = (I-\Gamma)^{\frac{1}{2}} \otimes (I-\Gamma)^{\frac{T}{2}} (I - \Gamma \otimes \Gamma^T)^{-1} \mathrm{vec}(\Sigma)$ with

$$\limsup_{n\to\infty} \sup_{\theta\in R_{n,0}} \left\{ \sigma_{min}\left(\Sigma_X\right)^{-1} + \sigma_{max}\left(\Sigma_X\right) \right\} < \infty.$$

Proof. For any $\theta \in \Theta$ we have

$$\begin{aligned} \left|\left| \mathbb{E}\left((I-\Gamma)^{\frac{1}{2}} X_{t,\theta} X_{t,\theta}^T (I-\Gamma)^{\frac{T}{2}} \right) \right|\right| &= \left|\left| (I-\Gamma)^{\frac{1}{2}} \sum_{s=0}^{t-1} \Gamma^s \Sigma \left(\Gamma^s\right)^T (I-\Gamma)^{\frac{T}{2}} \right|\right| \\ &\leq ||\Sigma|| \sum_{s=0}^{t-1} \left|\left| (I-\Gamma)^{\frac{1}{2}} \Gamma^s \right|\right|^2 \end{aligned}$$

Due to the block diagonal structure of Γ (Assumption U.4) and Assumption U.2, there exists some generic constant $c_0 > 0$ such the last term in above inequality is bounded by

$$c_0 \left(\sup_{\theta\in\Theta\,:\,|\lambda_N|>1-\alpha} \sum_{s=0}^{t-1} \left|\left| (I-\Gamma)^{\frac{1}{2}} \Gamma^s \right|\right|^2 + \sup_{\theta\in\Theta\,:\,|\lambda_1|\leq 1-\alpha} \sum_{s=0}^{t-1} \left|\left| (I-\Gamma)^{\frac{1}{2}} \Gamma^s \right|\right|^2 \right).$$

By equation (2.12.24), the second term is converges for $t \to \infty$. For the first term, we use the fact that, for any $\theta \in \Theta$, the condition $|\lambda_N| > 1 - \alpha$ implies that Γ is diagonal and, thus,

$$\sum_{s=0}^{t-1} \left\| (I-\Gamma)^{\frac{1}{2}} \Gamma^s \right\|^2 \leq \sum_{i=1}^{N} \sum_{s=0}^{t-1} m_i |1-\lambda_i| |\lambda_i|^s \leq d^2 r_\alpha \frac{1-|\lambda_i|^t}{|\lambda_i|} \leq \frac{d^2 r_\alpha}{1-\alpha}$$

where in the first inequality we used the fact that $|\lambda_i|^{2s} \leq |\lambda_i|^s$ since $|\lambda_i| \leq 1$ and in the second inequality we used Assumption U.3, the fact that $N, m_i \leq d$ and that $|\lambda_i| = 1$ implies that $\lambda_i = 1$ and therefore $|1-\lambda_i||\lambda_i|^s = 0$ for all $s = 0, \ldots, t-1$ in this case. This proves (2.K.3).

For the proof of (2.K.4), simply note that Lemma 2 and (2.K.3) imply that

$$\lim_{n\to\infty} \sup_{\theta \in R_{n,0}} \left|\left| (I-\Gamma)^{\frac{1}{2}} \left(H - \mathbb{E}(X_{n-1,\theta} X_{n-1,\theta}^T) \right) (I-\Gamma)^{\frac{T}{2}} \right|\right| = 0$$

and

$$\begin{aligned} &\sup_{\theta \in R_{n,0}} \left|\left| (I-\Gamma)^{\frac{1}{2}} \mathbb{E}(X_{n-1,\theta} X_{n-1,\theta}^T) (I-\Gamma)^{\frac{T}{2}} - \Sigma_X \right|\right| \\ = &\sup_{\theta \in R_{n,0}} \left|\left| (I-\Gamma)^{\frac{1}{2}} \sum_{s=n-1}^{\infty} \Gamma^s \Sigma \left(\Gamma^s\right)^T (I-\Gamma)^{\frac{T}{2}} \right|\right| \to 0 \end{aligned}$$

for $n \to \infty$. It remains to check that Σ_X is uniformly bounded and invertible in the limit. Since $\limsup_n \sup_{\theta \in R_{n,0}} \sigma_{max}(\Sigma_X) < \infty$ follows immediately from (2.K.3), we only need to show the latter. For any $\theta \in R_{n,0}$ diagonal, we have

$$\begin{aligned} \sigma_{min}(\Sigma_X) &\geq \sigma_{min}(\Sigma) \sum_{t=0}^{\infty} \sigma_{min} \left((I-\Gamma)^{\frac{1}{2}} \Gamma^t \right)^2 \\ &\geq \sigma_{min}(\Sigma) \sum_{t=0}^{\infty} \min_{1 \leq k \leq d} (1-|\lambda_{i_k}|) |\lambda_{i_k}|^{2t} \\ &\geq \sigma_{min}(\Sigma) \sum_{t=0}^{\infty} \frac{\log n}{n} \left(1 - \frac{\log n}{n}\right)^{2t} \\ &= \sigma_{min}(\Sigma) \frac{\log n}{n} \left(1 - \left(1 - \frac{\log n}{n}\right)^2\right)^{-1} \\ &\geq \frac{\sigma_{min}(\Sigma)}{2}. \end{aligned}$$

If Γ is non-diagonal, the same bound holds since adding ones on the super-diagonal does not decrease the minimum singular value. Because Σ is uniformly invertible over Θ, the proof is then complete. □

Proof of Lemma 2. Throughout the proof we write matrices as 3×3 block matrices such that the top-left block is $r_1 \times r_1$, the middle block is $(r_2-r_1)\times(r_2-r_1)$, and the bottom-left block is $(d-r_2)\times(d-r_2)$. We use a superscript to denote the block index, e.g., S_{ZZ}^{13} denotes the top-right block of S_{ZZ}. Furthermore, for a diagonal matrix A with complex values, we let $A^{\frac{1}{2}}$ denote the diagonal matrix obtained by taking the principal square root of the diagonal of A. Let $\tilde{Z}_t = \sum_{s=1}^{t}(1-n^{-\beta})^{t-s}\epsilon_s$ and $\psi_t = \sum_{s=1}^{t}(1-n^{-\beta})^{t-s}X_{s-1}$ so that

$$Z_t = \tilde{Z}_t + (\Gamma_n - I)\,\psi_t. \tag{2.K.5}$$

Let $\Lambda_n = (I-\Gamma_n)^{\frac{1}{2}}$ be the diagonal matrix as defined in Lemma 3 above. Then, since Γ_n and Λ_n commute, with c denoting some generic constant not depending on t or n,

$$\begin{aligned}
\mathbb{E}\,||\Lambda_n\psi_t||^2 &= \sum_{i,j=1}^{t}(1-n^{-\beta})^{2t-i-j}\mathrm{tr}\left(\Lambda_n\mathbb{E}\left(X_{i-1}X_{j-1}^T\right)\Lambda_n\right)\\
&\leq 2\sum_{1\leq j\leq i\leq t}(1-n^{-\beta})^{2t-i-j}\left|\mathrm{tr}\left(\Gamma_n^{i-j}\Lambda_n\mathbb{E}\left(X_{i-1}X_{j-1}^T\right)\Lambda_n\right)\right|\\
&\leq 2c\sum_{i,j=1}^{t}(1-n^{-\beta})^{2t-i-j}\left\|\Gamma^{i-j}\Lambda_n\right\|\\
&\leq 2c\sum_{i=0}^{t-1}(1-n^{-\beta})^{i}\sum_{j=0}^{t-i-1}\left\|\left((1-n^{-\beta})\Gamma\right)^{j}\Lambda_n\right\|
\end{aligned} \tag{2.K.6}$$

where the second inequality follows from Lemma 3 in the Appendix and the Cauchy-Schwartz inequality. This inequality yields a result equivalent to equation (40) in Phillips et al. [2009]. In particular, we deduce that $\sup_{1\leq t\leq n}\mathbb{E}\,||(\Gamma_n-I)\psi_t||^2 = o(n)$ from which it follows that

$$S_{Z\epsilon} = S_{\tilde{Z}\epsilon} + o_p(1). \tag{2.K.7}$$

We first prove (2.K.1). For ease of notation, we write $S_n = D_n^{\frac{1}{2}}S_{ZZ}D_n^{\frac{1}{2}}$. Since $D_n^{11} = n^{-\beta}I_{r_1}$, essentially the same proof as that of Lemma 3.1.(iii) in Phillips et al. [2009] using (2.K.6) shows that

$$S_n^{11} = n^{-\beta}S_{\tilde{Z}\tilde{Z}}^{11} + o_p(1) = \frac{1}{2}\Sigma^{11} + o_p(1),$$

where the latter equality follows from Lemma 2 and the fact that $n^{-\beta}\mathbb{E}(S_{\tilde{Z}\tilde{Z}}) \to \Sigma/2$ for $n\to\infty$. Similarly, the proof of Lemma 3.5.(ii) in Phillips et al. [2009] can be adapted to show that

$$S_n^{33} = \left(D_n^{33}\right)^{\frac{1}{2}}S_{XX}^{33}\left(D_n^{33}\right)^{\frac{1}{2}} + o_p(1) = \Sigma_{X,n}^{33} + o_p(1),$$

where the latter equality follows from Lemma 2 and Lemma 3 in the Appendix and $\Sigma_{X,n}^{33}$ is defined as in Lemma 3 in the Appendix but emphasizing the dependence on n. For the middle block, using the recursive relations $Z_t = (1-n^{-\beta})Z_{t-1} + \Delta X_t$ and $\Delta X_t = (\Gamma_n - I)X_{t-1} + \epsilon_t$, we can write

$$\left(1-\left(1-n^{-\beta}\right)^2\right)S_{ZZ}^{22} = S_{\Delta XZ}^{22} + S_{Z\Delta X}^{22} + S_{\Delta X\Delta X}^{22} + o_p(1).$$

It follows from Lemma 2, that $S^{22}_{\Delta X \Delta X} = \Sigma^{22} + o_p(1)$. For the other two terms, we use (2.K.7) and write $S^{22}_{Z\Delta X} = S^{22}_{\tilde{Z}\epsilon} + S^{22}_{ZX}(\Gamma^{22}_n - I)^T + o_p(1)$. Using the recursive relations and (2.K.7) once more yields

$$\left(I - \left(1 - n^{-\beta}\right)\Gamma^{22}_n\right) S^{22}_{XZ} = S^{22}_{X\epsilon} + S^{22}_{\epsilon\tilde{Z}} + S^{22}_{\epsilon\epsilon} + S^{22}_{XX}(\Gamma^{22}_n - I) + o_p(1)$$

It follows from Lemma 2 that the first two terms tend to 0 in probability for $n \to \infty$ and $S^{22}_{\epsilon\epsilon} = \Sigma^{22} + o_p(1)$. If we define $K \in \mathbb{C}^{(r_2-r_1)\times(r_2-r_1)}$ diagonal with $K_{ii} = \kappa_i$ for $i \leq r_2 - r$ and $K_{ii} = \kappa_i^{-1}$ otherwise, we get $n^\beta\left(I - \Gamma^{22}_n\right) \to K$, $n^\beta \Lambda^{22}_n \to K^{\frac{1}{2}}$, and $n^\beta\left(I - (1 - n^{-\beta})\Gamma^{22}_n\right) \to K + I$ for $n \to \infty$. Lemma 3 in the Appendix then yields

$$\left(\Gamma^{22}_n - I\right) S^{22}_{XZ} = (K + I)^{-1}\left(K\Sigma^{22} + K^{\frac{1}{2}}\Sigma^{22}_{X,n}K^{\frac{1}{2}}\right) + o_p(1).$$

and $(\Lambda^{22}_n)^{\frac{1}{2}} \otimes (\Lambda^{22}_n)^{\frac{T}{2}}(I - \Gamma^{22}_n \otimes (\Gamma^{22}_n)^T)^{-1} \to (K \otimes K^T)^{\frac{1}{2}}(I \otimes K^T + K \otimes I)^{-1}$ for $n \to \infty$ so that (noting that K_{ii} has strictly positive real part for all i)

$$\Sigma^{22}_{X,n} \to K^{\frac{1}{2}} \int_0^\infty e^{-sK}\Sigma^{22}e^{-sK^T} ds K^{\frac{T}{2}} = K^{\frac{1}{2}}\Omega^{22}K^{\frac{T}{2}}.$$

Then, using the relation $\Sigma^{22} - \Omega^{22}K = K\Omega^{22}$, the limiting expression simplifies to

$$(\Gamma^{22}_n - I)S^{22}_{XZ} = (K + I)^{-1}K^2\Sigma^{22}.$$

Finally, since $n^\beta(1 - (1 - n^{-\beta})^2) = 2 + o(1)$ and $D^{22}_n = n^{-\beta}I_{r_2-r_1}$, we find

$$\begin{aligned} S^{22}_n &= \frac{1}{2}\left(\Sigma^{22} + (K + I)^{-1}K^2\Omega^{22} + \Omega^{22}(K^T)^2(K + I)^{-T}\right) + o_p(1) \\ &= \frac{1}{2}\left((K + I)^{-1}K\Omega^{22} + \Omega^{22}K^T(K + I)^{-T}\right) + o_p(1) \\ &= \frac{1}{2}(I + K)^{-1}\left(2K\Omega^{22}K^T + \Sigma\right)(I + K)^{-T} + o_p(1). \end{aligned}$$

We have yet to characterize the asymptotic behaviour of the off-diagonal blocks. First, note that by (23) in Phillips et al. [2009] and (2.K.5), we get

$$S^{32}_{ZZ} - S^{32}_{XZ} = -n^{-\beta}\left(S^{32}_{\psi\psi}(\Gamma^{22}_n - I)^T - S^{32}_{\psi\tilde{Z}}\right)$$

so that (2.K.6) and Lemma 2 yield $S^{32}_n - \left(D^{33}_n\right)^{\frac{1}{2}} S^{32}_{XZ}\left(D^{22}_n\right)^{\frac{1}{2}} = o_p(1)$. Similar to above, we have

$$\left(I - (1 - n^{-\beta})\Gamma^{33}_n\right) S^{32}_{XZ} = S^{32}_{XX}\left(\Gamma^{22}_n - I\right)^T + o_p(1).$$

But then, because $n^{-\beta}(I - \Gamma^{33}_n) = o(1)$, we find that

$$n^{-\frac{\beta}{2}}\left(\Lambda^{33}_n\right)^{\frac{1}{2}}\left(I - \left(1 - n^{-\beta}\right)\Gamma^{33}_n\right)^{-1} = o(1)$$

and, by Lemma 3 in the Appendix and Lemma 2, $S^{32}_{XX}(\Gamma^{22}_n - I)^T = o_p(1)$. Thus, $S^{32}_n = o_p(1)$. A similar argument show that $S^{31}_n = o_p(1)$ so that the only block left is S^{21}_n. As a consequence of (2.K.5), we have

$$n^{-\beta}\left|\left|S^{12}_{ZZ} - S^{12}_{\tilde{Z}Z}\right|\right| = n^{-\beta}\left|\left|(\Gamma^{11}_n - I)S^{12}_{\psi\psi}(\Gamma^{22}_n - I)^T + (\Gamma^{11}_n - I)S^{12}_{\psi\tilde{Z}}\right|\right|$$

and arguments like the one employed in the proof of Lemma 3.1 in Phillips et al. [2009] in combination with (2.K.6) shows that the right hand side is $o_p(1)$. Using the recursive relations for Z_t, $\tilde{Z}_t$, and ΔX_t in combination with (2.K.7) and Lemma 2, we have

$$\begin{aligned}\left(1 - (1 - n^{-\beta})^2\right) S^{12}_{\tilde{Z}Z} &= S^{12}_{\epsilon Z} + S^{12}_{\tilde{Z}\Delta X} + S^{12}_{\epsilon\Delta X} + o_p(1) \\ &= S^{12}_{\tilde{Z}X}(\Gamma^{22}_n - I)^T + S^{12}_{\epsilon\epsilon} + o_p(1).\end{aligned}$$

An application of Lemma 3 in the Appendix and Lemma 2 yields $S^{12}_{\tilde{Z}X} = \Omega^{12}_n K^{\frac{T}{2}} + o_p(1)$ where

$$\Omega^{12}_n = \Sigma^{12}\left(I - (1 - n^{-\beta})\Gamma^{22}_n\right)^T n^{-\frac{\beta}{2}}\left(I - \Lambda^{22}_n\right)^{-\frac{T}{2}} \to \Sigma^{12}K^{\frac{T}{2}}(I + K)^{-T}$$

for $n \to \infty$. In conclusion, since $D^{22}_n = n^{-\beta}I_{r_2 - r_1}$ and $D^{33}_n = n^{-\beta}I_{r_1}$, we get

$$S^{12}_n = \frac{1}{2}\left(\Sigma^{12} + \Sigma^{12}K^T(I + K)^{-T}\right) + o_p(1) = \frac{1}{2}\Sigma^{12} + o_p(1).$$

Collecting all the limiting expressions, we define

$$\bar{K} = \begin{pmatrix} I_{r_1} & 0 \\ 0 & K \end{pmatrix}, \quad \Omega = \int_0^\infty e^{-s\bar{K}}\begin{pmatrix}\Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22}\end{pmatrix}e^{-s\bar{K}^T}ds$$

and observe that

$$\begin{aligned}\Sigma^{11} &= \left((I + \bar{K})^{-1}\left(2\bar{K}\Omega\bar{K}^T + \Sigma\right)(I + \bar{K})^{-T}\right)^{11} \\ \Sigma^{12} &= \left((I + \bar{K})^{-1}\left(2\bar{K}\Omega\bar{K}^T + \Sigma\right)(I + \bar{K})^{-T}\right)^{12}.\end{aligned}$$

so that $S_n = \Sigma_{Z,n} + o_p(1)$ with

$$\Sigma_{Z,n} = \frac{1}{2}\begin{pmatrix}(I + \bar{K})^{-1}\left(2\bar{K}\Omega\bar{K}^T + \Sigma\right)(I + \bar{K})^{-T} & 0 \\ 0 & \Sigma^{33}_{X,n}\end{pmatrix}.$$

To see that $\Sigma_{Z,n}$ is asymptotically invertible and bounded simply note that the real part of $\bar{K}_{ii}$ is in $[0, 1]$ for all $1 \leq i \leq r_2$ and Σ is positive definite. Therefore, the top left block of $\Sigma_{Z,n}$ is some fixed positive definite matrix for all $n \in \mathbb{N}$. The result then follows from Lemma 3 in the Appendix.

Once (2.K.1) has been established, the proof of (2.K.2) is completely analogous to the proof of equation (2.10.8). □

Proof of Theorem 6. It follows from Lemma 1 and 2 in combination with Proposition 8 in the supplementary material for Lundborg et al. [2022] that $\hat{t}^2_{IV,\Gamma} \to_w \chi^2_{d^2}$ uniformly over Θ. □

3 Beyond stationarity: Nonlinear cointegration

This chapter contains the following paper:

[**STEM**] [Holberg, 2024]. C. Holberg. Stationary embeddings: A nonlinear generalization of cointegration, 2024. Working paper.

The paper is still only a manuscript. We emphasize that the results are preliminary and mistakes might appear. Our main purpose is to introduce a sensible nonlinear generalization of cointegration. We take as our starting point the blind source separation formulation of cointegration. In other words, given an observable process x_t, we assume that we can write $x_t = d(y_t, z_t)$ where $y_t \in \mathbb{R}^k$ is the stationary latent component, $z_t \in \mathbb{R}^{p-k}$ is the non-stationary latent component, and $d : \mathbb{R}^p \to \mathbb{R}^p$ is some invertible smooth mixing. The goal is then to identify and estimate the *stationary embedding* $e = d^{-1}$ that maps x_t back into its constituent parts. We discuss some applications at the end of the paper. Notably, under suitable conditions, this definition of nonlinear cointegration is closely related to manifold learning.

Stationary Embeddings: A Nonlinear Generalization of Cointegration

Christian Holberg

Abstract

Most signals arising in practice exhibit non-stationary behavior, but in many cases such signals can be decoupled into a stationary and non-stationary component via a smooth invertible transformation. We discuss in what sense these component processes are identifiable and provide a method for estimating the decoupling transformation based on signature kernels. The resulting framework can be seen as a nonlinear generalization of cointegration. Finally, we discuss some important applications of the developed methodology and, in particular, how it relates to manifold learning.

3.1 Introduction

A stochastic process is stationary if its law does not depend on time. This is in many ways a desirable property. For one, it ensures that the future behaves more or less like the present so that, intuitively, any inference we might make on the data we have at hand today can be expected to hold also tomorrow. Another point, of a more technical nature, is that a lot of the asymptotic theory for i.i.d. data carries over to the stationary setting — sometimes under additional mixing and ergodicity assumptions. This enables us to adapt many of the existing methods that are already prevalent in statistics and machine learning. However, whatever preferences we may have, the so-called real world need not conform to them. As is obvious to anyone who has spent a lot of time wrangling with time-series data, it is rarely the case that it can be said to be stationary. At least not without any further pre-processing. And why should it? Stationarity, for example, precludes any trend in the mean or spread of the process. It also does not allow for a data generating process whose distribution changes over different environments. All of these are common characteristics of observed stochastic processes.

There is a sense in which non-stationary processes are actually the ones of interest in that this is where the change happens. To illustrate this point, consider a process x_t that can be written as the sum of a stationary component, y_t, and a non-stationary component, z_t: $x_t = z_t + y_t$. If the mean of z_t changes greatly over time (compared to the variance of y_t) or if it behaves like a martingale so that its variance increases over time,

then it is clear that the behavior of x_t is mostly determined by z_t and y_t can be viewed as something like an error term. The focus of the present paper is exactly processes such as x_t. To be more precise, we consider settings in which one observes a non-stationary process x_t of some dimension $p \geq 2$ that can be written as a mixture of two unobserved processes $y_t \in \mathbb{R}^k$ and $z_t \in \mathbb{R}^{p-k}$ where the one is stationary and the other not, that is, we have $x_t = d(y_t, z_t)$ for some smooth invertible map $d : \mathbb{R}^p \to \mathbb{R}^p$. It turns out that, under a certain Hölder-continuity assumption on the mixing transformation, d, x_t will stay close to a $(p-k)$-dimensional manifold with deviations governed by the stationary process y_t (see Proposition 1). If the variance of y_t is not too large, learning this manifold could prove advantageous and can be viewed as a dimensionality reduction technique. On the other hand, there are many settings where we would instead be interested in the stationary component. Indeed, before doing any analysis, the statistician will often start by removing trends, seasonality, etc. so that the resulting process is stationary. Our method can then be seen as a general non-parametric way of retrieving the stationary component. Another example is if we wish to extrapolate beyond the observed time frame; something that is especially relevant for classification or regression problems where learning is done on a small initial window and prediction happens long after. We return to these applications in more detail at the end of the paper.

In essence, given an observation of x_t, our objective then lies in learning the inverse mixing transformation $e := d^{-1}$ that decomposes x_t back into its two latent components. We also call this map the *stationary embedding*. Readers familiar with cointegration will realize the similarities. Indeed, if z_t is integrated of order 1[1] with no linear combination of its coordinates being stationary and d is an invertible linear transformation, this corresponds exactly to the usual definition of cointegration (see, e.g., Engle and Granger [1987]). Importantly, in the linear setting, the order of integration of x_t corresponds to that of the non-stationary component. This is no longer the case when d can be any arbitrary transformation. Similarly, one usually only cares about *weak* stationarity of the process y_t which, of course, is also not preserved by most non-linear transformations (if y_t is non-stationary). Thus, in order to extend cointegration to the nonlinear setting, we shall work with the strict definition of stationarity and consider general deviations from it. Instead of assuming that all linear combinations of z_t are integrated of order 1, we shall assume that all invertible smooth transformations of z_t are non-stationary (see our Definition 4).

Two main difficulties arise as we try to define nonlinear cointegration. The first one concerns identifiability. In the linear case, it is well-known that only the stationary component is identifiable (see also Von Bünau et al. [2009] for the same result but under the framing of *stationary subspace analysis*). Extending this result to our case would then be equivalent to saying that we can identify y_t up to smooth invertible transformations. Once we have established what exactly is identifiable, the second difficulty we need to resolve is then how one goes about estimating the inverse mixing, e. This time

[1] In the context of the present paper this shall simply mean that z_t is non-stationary with stationary first differences, i.e., $z_{t+\delta} - z_t$ is stationary for all $\delta > 0$. One can keep in mind a Levy process as the canonical example.

we cannot draw inspiration from the linear case where one usually relies on the fact that the empirical covariance matrix of the non-stationary component diverges as the sample size increases (for a non-parametric method relying rather directly on this fact, we refer to Zhang et al. [2019], but even the methods of Johansen [1988, 1991, 1995] and Phillips [1991] would not be able to separate the stationary and non-stationary parts otherwise). Instead we need a general way to distinguish stationary from non-stationary processes. We do this by using feature map stemming from rough path theory known as the *signature*. In particular, similar to Issa et al. [2023a], we measure the stationarity of a process by comparing the distribution of different slices of its path using the signature kernel maximum mean discrepancy.

3.1.1 Contributions

The main contributions of our work are the following. First, we establish clear conditions on either the class of admissible mixings, d, or on the kind of non-stationarity of z_t that allow for identification results akin to those that hold in the linear case. In particular, we find that, under either of these conditions, the stationary latent process, y_t, is identifiable up to smooth invertible transformations. The main result is encapsulated in Theorem 1. Along the way we also show that a wide class of multivariate processes exist that are non-stationary under general invertible smooth transformations. Although all of these results are stated for continuous time processes, they carry over to discrete time with no major modifications.

Our second contribution lies in defining a statistic that allows for estimating the inverse mixing. In the ideal setting where we have access to the underlying distribution, this statistic is 0 if, and only if, the process is stationary. In practice one usually only observes a single long discretely sub-sampled trajectory. We define an approximate objective and establish under what conditions this objective yields consistent inference. This is the statement of Theorem 3.

Finally, we develop feasible algorithms for computing the approximate objective (i.e., Algorithm 3 and 4) and thus estimating the stationary embedding up to its equivalence class. We give a range of applications at the end of the paper one of them being manifold learning for non-stationary stochastic processes.

3.1.2 Related work

Nonlinear ICA

Cointegration can be seen as a form of *blind source separation* (BSS) for which a rich literature exists [Choi et al., 2005]. A particularly important strand is called *independent component analysis* (ICA) where it is assumed that the observed process is an invertible mixing of latent independent processes. Identifiability has been established under general conditions [Hyvarinen et al., 2019, Hyvarinen and Morioka, 2017, Hyvarinen et al., 2019, Schell and Oberhauser, 2023] and for different variants of the problem Khemakhem et al. [2020a,b]. In particular, our definition of identifiability is close to the one of *block-identifiability* given in Von Kügelgen et al. [2021]. A common way to estimate

the de-mixing transformation utilizes contrastive learning approaches [Hyvarinen and Morioka, 2016]. Our approach is more akin to that of [Schell and Oberhauser, 2023] who show that the de-mixing transformation can be found by minimization of a function characterizing independence.

Cointegration

By now cointegration is a very mature field. It started with the seminal work in Engle and Granger [1987] with much of the theory developed in Johansen [1988, 1991, 1995] for the vector autoregressive model and in Phillips [1991] in a more general setup. There is also a part of the literature where many of the same ideas are explored under the nomenclature of stationary subspace analysis (SSA) [Von Bünau et al., 2009, Sundararajan and Pourahmadi, 2018, Baktashmotlagh et al., 2014]. Quite a few attempts in the direction of nonlinear cointegration have been made either by extending the vector error correction model [Escribano, 2004, Kristensen and Rahbek, 2007, Balke and Fomby, 1997], in the form of nonlinear regression with non-stationary predictor and target [Park and Phillips, 1999, 2001], or as non-stationary non-linear autoregressive Markov chains [Karlsen and Tjøstheim, 2001, Karlsen et al., 2007, Li et al., 2016]. A recent review can be found in Tjøstheim [2020]. Recently Duffy et al. [2022] presented a way in which nonlinear cointegration relationships can arise in an autoregressive time series model under a very specific type of nonlinearity. None of these works are fully satisfactory though because they do not treat the problem of finding fully general nonlinear cointegration relations.

Change point detection

Change point detection methods also deal with non-stationary processes, although the goal is slightly different. Usually one assumes that the process is piece-wise stationary. The aim is then to find the points at which the distribution changes given a sampled trajectory (offline detection) [Truong et al., 2020] or to detect a change point as the data arrives (online detection) [Aminikhanghahi and Cook, 2017, Adams and MacKay, 2007]. Our statistic discriminating stationary from non-stationary processes (see Section 3.3) is similar in spirit to some of the kernel change point detection methods [Li et al., 2015, Garreau and Arlot, 2018, Harchaoui et al., 2008]. In Issa and Horvath [2023] the authors develop a method for online market regime detection based on the signature kernel maximum mean discrepancy. Their test statistic is quite similar to the statistic for determining stationarity discussed here. There are, however, two key differences. First, they partition their data into so-called ensemble paths which is similar to our batching protocol. Whereas we allow for very general partitions, in Issa and Horvath [2023] the authors restrict themselves to sliding disjoint windows. The other big difference is that we employ the discrete signature and, in particular, the Fourier approximation of Toth et al. [2023]. Our main reason for doing so is that we need to evaluate our statistic as part of an optimization problem necessitating computational efficiency. Efficient algorithms are developed in Appendix 3.D.

3.1.3 Notation

In this section we collect some notation that reappears throughout the thesis. Some of the following will be reintroduced in the relevant places below, but use this section as a glossary of sorts.

We use small letters to denote stochastic process. If confusion can be avoided, we will often simply write x_t when referring to the whole stochastic process $(x_t)_{t\in\mathcal{T}}$. As far as possible, we will be consistent and use $\mathcal{T}$ for the index set. $\mathcal{T}$ is then either an interval of the form $[s,t]$ or $[s,\infty)$ (usually with $s=0$) or it will be a discrete (possibly infinite) set of points $\mathcal{T}=(t_0<t_1<\cdots<t_n)$ also known as a *time grid*. We generally reserve capital letters for Euclidean random variables. Thus, a discrete stochastic process, or time series, of the form $(x_t)_{t\in\mathcal{T}}$ with $\mathcal{T}=(t_0<t_1<\dots)$ may equivalently be represented as a sequence of random variables $X_1,X_2,\dots$ where $X_k=x_{t_k}$. Overloading the notation, we use Δ for two purposes: 1) denoting the first difference of a time series, $\Delta x_{t_k}=x_{t_k}-x_{t_{k-1}}$, and 2) for the set of ordered k-tuples over some set, that is, $\Delta_k(I):=\{(t_1,\dots,t_k)\in I^k \mid t_1<\cdots<t_k\}$ where $I\subset\mathbb{R}$. For any $n\in\mathbb{N}$ we define $[n]:=\{1,\dots,n\}$. Equality in law is expressed using $=_d$. We reserve π for projections adding a subscript to clarify the co-domain. For a p-dimensional vector space, V, and $1\le i\le j\le p$, we use $\pi_{i:j}$ to denote the projection onto the subspace spanned by the basis vectors $e_i,\dots,e_j$ of V (where we assume some standard fixed basis is in use). For example, if $V=\mathbb{R}^p$, then $\pi_{i:j}$ corresponds to the projection onto the coordinates between i and j. Finally, we reserve θ for the shift operator. That is, for $s\in\mathbb{R}_+$ and some vector space V, we define $\theta_s:\mathbb{R}_+^V\to\mathbb{R}_+^V$ given by $\theta_s f(\cdot)=f(s+\cdot)$. The shift operator is defined analogously for sequences.

3.2 Stationary embeddings

This section lays the theoretical groundwork for the rest of the article. We will spend most of it clarifying (in a mathematically rigorous way) what we mean by non-linear cointegration and stationary embeddings. In the following we shall use the term cointegration for both the linear and non-linear case. We acknowledge that this is an abuse of terminology in the sense that, for the linear case, cointegration very specifically refers to integrated processes that upon applying specific linear transformations are integrated of a lower order. Here we are concerned with general non-stationary processes and we are simply interested in finding a transformation (linear or not) taking this process to a lower-dimensional stationary process. As such, our concept of cointegration, arguably, is more akin to that considered in linear stationary subspace analysis (SSA) [Von Bünau et al., 2009, Sundararajan and Pourahmadi, 2018, Baktashmotlagh et al., 2014]. However, the definition of cointegration seems to vary considerably in the existing literature and it is not entirely clear what separates it from SSA apart from the algorithms employed.

First, let us define what exactly is meant by a *stationary* process. For an interval $J\subset\mathbb{R}_+$, we let $C(J,\mathbb{R}^p)$ denote the space of continuous functions on J taking values in $\mathbb{R}^p$ omitting the domain when it is clear from context. For some topological space $\mathcal{X}$, we

define $\mathcal{M}_1(\mathcal{X})$ as the set of probability measures on $\mathcal{X}$ (equipped with its Borel sigma-algebra). We define the shift operator $\theta_\tau : C(\mathbb{R}^p) \to C(\mathbb{R}^p)$ such that $\theta_\tau : x_\cdot \mapsto x_{\tau+\cdot}$. Finally, for any $\mu \in \mathcal{M}_1(C(\mathbb{R}^p))$ and $0 \leq s \leq t$, we define $\mu_{[s,t]} := \mu \circ \pi_{[s,t]}^{-1}$.

Definition 1. For any $t > 0$ and $p \geq 1$, we define the set of probability measures

$$\mathcal{S}^p(t) := \left\{ \mu \in \mathcal{M}_1(C([0,\infty)), \mathbb{R}^p) \mid \mu_{[0,t]} = \mu_{[s,s+t]} \text{ for all } 0 \leq s \right\}.$$

We also define $\mathcal{S}^p := \bigcap_{t \geq 0} \mathcal{S}^p(t)$. Any process $x \sim \mu$ with $\mu \in \mathcal{S}^p$ is *stationary*. We shall often abuse notation and simply write $x \in \mathcal{S}^p(t)$ or $x \in \mathcal{S}^p$.

The definition of stationary processes given above agrees with the one involving finite-dimensional distributions. Indeed, if $x \in \mathcal{S}^p$, then, for any $n \geq 1$ and $0 \leq t_1, \ldots, t_n, \tau$, we have $x \in \mathcal{S}^p(t_n)$ which implies that the distributions of $(x_{t_1}, \ldots, x_{t_n})$ and $(x_{t_1+\tau}, \ldots, x_{t_n+\tau})$ coincide. For ease of notation, we define $J_T = [0, T]$ if $T < \infty$ and $J_\infty = \mathbb{R}_+$. For a function $f : \mathbb{R}^p \to \mathbb{R}^q$ and some integers $1 \leq i \leq j \leq q$ we define $f_{i:j} = \pi_{i:j} \circ f : \mathbb{R}^p \to \mathbb{R}^{j-i+1}$. We also define the spatial support of a stochastic process. We shall assume through out that the spatial support of the observed process x, call it D_x, is path-connected and convex.

Definition 2 (Spatial support)**.** For a stochastic process $w \in C(J_T, \mathbb{R}^q)$ we define the *spatial support* of w to be the set

$$D_w = \overline{\bigcup_{t \in J_T} \operatorname{supp}(w_t)}$$

where supp denotes the usual support of a Euclidean random variable.

Before we give the main definition of this section, namely, that of *stationary embeddings* and *cointegration*, let us first recall the linear case. Loosely speaking, for a non-stationary process $x_t \in \mathbb{R}^p$ we say that it is cointegrated of order $0 \leq k \leq p$ if there exists an invertible matrix $Q \in \mathbb{R}^{p \times p}$ such that x_t can be decomposed $Qx_t = (y_t, z_t)$ with $y_t \in \mathbb{R}^k$ stationary and $z_t \in \mathbb{R}^{p-k}$ non-stationary and k being the largest number for which such a decomposition is possible. In other words, $x_t = Q^{-1}(y_t, z_t)$ is a mixture of two latent components distinguished by their degree of stationarity. Now, a natural way to generalize this definition is to replace the mixing matrix Q^{-1} by a general diffeomorphism. As we shall see, however, this function class is too big to obtain meaningful identification results without any further assumptions on the underlying latent non-stationary component. In the following we allow for different classes of mixing transformations depending on which assumptions we put on z_t. There are, however, some common properties which any potential candidate class of mixing transformations should satisfy. In particular, we want them to be invertible and closed under composition and inversion.

Definition 3 (Admissible mixing)**.** Let $\mathcal{D}$ be the class of piece-wise[2] diffeomorphic homeomorphisms and $\mathcal{F}$ a function class (see Def. 10). We say that $\mathcal{F}$ *is a class of*

[2]We refer to Def. 11 for a precise definition of what it means for a function to be of property $\mathcal{P}$ piece-wise.

admissible mixings if $\mathcal{F} \subset \mathcal{D}$ and, for any $f \in \mathcal{F}(U)$ and $g \in \mathcal{F}(f(U))$, it holds that f is invertible with $f^{-1} \in \mathcal{F}(f(U))$ and $g \circ f \in \mathcal{F}(U)$.

Notice that in our definition of cointegraiton we require that the non-stationary component z_t is not itself cointegrated. In order to obtain a sort of minimal (in terms of the cointegration rank k) decomposition we shall therefore define a process to be *strictly non-stationary* if it is not a mixing of a stationary and non-stationary component. This is exactly the content of the next definition albeit phrased slightly differently.

Definition 4 (Strictly non-stationary)**.** We say that a process $w \in C(\mathbb{R}^q)$ is *strictly non-stationary* if there exists an open superset $U \supset D_w$ such that the coordinate processes of $f(w_t)$ are non-stationary for every $f \in \mathcal{D}(U)$.

An important question to ask oneself is then how big this class of processes actually is. In order to obtain any meaningfully general notion of non-linear cointegration, we should be able to show that many of the non-stationary processes used in practice are actually strictly non-stationary. This would follow as an easy corollary to Conjecture 1 for which we unfortunately do not yet have a proof.

We are ready to introduce the main definition of this paper, namely, a nonlinear generalization of cointegration or, equivalently, what we call *stationary embeddings*.[3]

Definition 5 (Stationary embedding)**.** Given a class of admissible mixings, $\mathcal{F}$, we call a map e a stationary embedding of $x \in C(\mathbb{R}^p)$ of dimension $0 \leq k \leq p$ if there is some open $U \supset D_x$ such that $e \in \mathcal{F}(U)$ and k is the largest integer such that $y_t := e_{1:k}(x_t) \in \mathcal{S}^k$.

We say that x_t is *cointegrated of order* k if it has a stationary embedding of dimension k, and we call x_t *strictly cointegrated of order* k (or $\mathcal{CI}_k$) if it is cointegrated of order k and $e_{(k+1):p}(x_t)$ is strictly non-stationary for every stationary embedding of dimension k.

Remark 1. The definition of strict cointegration might seem a little strange at first glance. In particular, it is not clear whether a process can be strictly cointegrated of different orders, say $0 \leq k < l \leq p$. We would really prefer for this not to be possible. Indeed, our goal with defining the concept of *strict* cointegration was to achieve a minimal decomposition, i.e., if x_t is strictly cointegrated of order k, then k should be the smallest dimension for which a stationary embedding exists. Luckily, for all relevant cases, namely the ones where the stationary embedding is *identifiable* (see also Def. 6), there is only one $0 \leq k \leq p$ for which x is $\mathcal{CI}_k$. We call this number the *cointegration rank* of x_t.

For a stationary embedding e of dimension k we shall write $e = (e_1, e_2)$ with $e_1 : \mathbb{R}^p \to \mathbb{R}^k$ and $e_2 : \mathbb{R}^p \to \mathbb{R}^{p-k}$. Comparing with linear cointegration, $e_1(x_t)$ are the cointegration relations and $e_2(x_t)$ correspond to the shared stochastic trends. Also, with $y_t := e_1(x_t)$ and $z_t := e_2(x_t)$, we find that $x = d(y_t, z_t)$ with $d := e^{-1}$ an admissible mixing on some open superset $U \supset D_{(y,z)}$, i.e., requiring that e be admissible ensures that we can write x_t as the mixture of a stationary and non-stationary component under an admissible mixing.

[3]The terminology is perhaps a little misleading. However, in a certain sense, as is made more precise in Proposition 1, the map d indeed embeds a k-dimensional manifold in $\mathbb{R}^p$ playing a similar role as invariant manifolds do for deterministic dynamical systems.

3.2.1 Identifiablity

Let $x_t \in \mathbb{R}^p$ be $\mathcal{CI}_k$ and write $x_t = d(y_t, z_t)$. The problem of finding y_t and z_t (or, equivalently, the stationary embedding, e) from an observation of x_t can be viewed as a blind source separation (BSS) problem [Choi et al., 2005]. At this point it is then worth pondering to what extent the components are identifiable given that the only thing available to use is the mixed process x_t. It is well-known that in the linear case, without any further assumptions on the underlying processes, the stationary component can only be identified up to invertible linear transformations and the non-stationary component remains unidentified. This is the same as saying that the non-stationary subspace is identified while the stationary subspace is unidentified [Von Bünau et al., 2009]. Below we find that a similar statement holds in the general non-linear setting. The main result is Theorem 1.

The way in which we state our identifiability results may seem indirect. Our aim is to end up at a result similar to the linear case; We want to be able to identify the stationary transformation up to invertible transformations, that is. It is clear that there are two levers to pull. On one hand, we can restrict the function class of admissible mixings where the smallest extreme is that of analytic functions. On the other hand, we can impose further assumptions on the latent non-stationary component, in particular, the degree to which the non-stationary behavior happens locally (in the support) or globally. We shall therefore first define what we exactly mean by identifiability and subsequently give examples of pairs of function classes and non-stationary processes which yield identifiability of the stationary embedding.

Definition 6. Let $\mathcal{F}$ be a function class of admissible mixings and $z \in C(\mathbb{R}^p)$.[4] We say that the pair $(\mathcal{F}, z)$ is *identifiable* if the following holds: Let $k \in [p]$ and $x_t \in C(\mathbb{R}^p)$ be $\mathcal{CI}_k$ with $e \in \mathcal{F}(U)$ a stationary embedding and $(y_t, \pi_{(k+1):p}(z_t)) = e(x_t)$. For any $f \in \mathcal{F}(U)$, we have

$$\tilde{y}_t := f_{1:k}(x_t) \in \mathcal{S}^k \quad \text{if and only if} \quad \tilde{y}_t = h(y_t) \text{ for some } h \in \mathcal{F}(e_1(U)). \tag{3.2.1}$$

Remark 2. This definition of identifiability is a natural generalization of the linear case. Indeed, if we let the class of admissible mixings be given by all invertible linear transformations and z is a strictly second order non-stationary process, then Equation (3.2.1) is exactly the well-known identification result from SSA. We also note that, by definition, classes of admissible mixings are closed under inversion. Thus, we could just as well have required $e^{-1} := d \in \mathcal{F}(e(U))$.

In words, Definition 6 states that a pair $(\mathcal{F}, z)$ is identifiable if, given the observation of any admissible mixing $x_t = d(y_t, \pi_{(k+1):p}(z_t))$, we are able to recover the stationary embedding up to homeomorphisms in the function class $\mathcal{F}$. We may define the equivalence relation $\sim_{\mathcal{F}}$ over $\mathcal{F}$ where $(U, f) \sim_{\mathcal{F}} (V, g)$ if and only if $g = h \circ f$ and $V = h \circ f(U)$ for some $h \in \mathcal{F}(f(U))$. For the most part, we will admit a slight abuse of notation and

[4] The reader may be confused why we are using p to denote the dimension of the non-stationary component all of a sudden contrary to $p - k$ as above. We do this to allow the same definition of identifiability across all cointegration ranks.

simply write $f \sim_{\mathcal{F}} g$. With e denoting the underlying stationary embedding $e := d^{-1}$, we can then define the equivalence class $[e]_{\mathcal{F}}$ given by all $f \in \mathcal{F}(U)$ such that $f_{1:k} \sim_{\mathcal{F}} e_1$. What we seek is the ability to be able to identify $[e]_{\mathcal{F}}$ based on an observation of x_t alone.

Before we state the main theorem, we introduce a regularity assumption. In particular, similar to Schell and Oberhauser [2023], we shall require that the latent non-stationary component is sufficiently noisy so as to admit a continuous density wrt. Lebesgue measure for any colleciton of time points.

Definition 7 (Regular process)**.** Let $z \in C(\mathbb{R}^p)$ with spatial support D_z. We say that z is *regular* if, for every $m \geq 1$ and $\mathbf{t} \in \Delta_m([0, \infty))$, the random variable $z_{\mathbf{t}} = (z_{t_1}, \ldots, z_{t_m})$ admits a continuous density with respect to pm-dimenisonal Lebesgue measure.

We now give two canonical examples of identifiable pairs $(\mathcal{F}, z)$. First, if we let $\mathcal{F}$ be the class of analytic functions with invertible Jacobian, then any strictly non-stationary and regular process will do. On the other end of the spectrum, if we assume the kind of non-stationarity of z_t to be *global* in a certain sense, we can work with all of $\mathcal{D}$, i.e., the class of piece-wise diffeomorphic homeomorphisms. Since the proof in both cases is almost identical we provide both results as one theorem.

Theorem 1 (Identifiability)**.** *Let $z \in C(\mathbb{R}^p)$ be strictly non-stationary and regular and $\mathcal{F}$ a class of admissible mixings. Then, the pair $(\mathcal{F}, z)$ is identifiable in the following two cases:*

(i) $\mathcal{F}$ only contains analytic functions.

(ii) The law of z_t is strictly non-stationary when restricted to U for any open $U \subset D_z$.

Proof. Any one-to-one transformation of a stationary process is still stationary. It therefore suffices to prove the "only if"-direction in (3.2.1). So suppose that $\tilde{y}_t$ is stationary and define $g := f_{1:k} \circ d : e(U) \to \mathbb{R}^k$ where $f : \mathcal{F}(U)$, $U \supset D_x$ is open and e is the stationary embedding of x_t, i.e., $d = e^{-1}$. We will prove the slightly stronger assertion

$$\partial_v g(u, v) = 0 \text{ for all } (u, v) \in V \text{ and some some open dense } V \subset e(U) \tag{3.2.2}$$

from which the result then readily follows by continuity.

Part (i): Since z_t is regular, we can find some open $V \subset e(U)$, time points $0 \leq t_1 \leq \cdots \leq t_n$, and a shift $\tau > 0$ such that, for any measurable $V' \subset V$

$$\mathbb{P}\left((y_{t_1}, z_{t_1}) \in V', \ldots, (y_{t_n}, z_{t_n}) \in V'\right) \neq \mathbb{P}\left((y_{\tau+t_1}, z_{\tau+t_1}) \in V', \ldots, (y_{\tau+t_n}, z_{\tau+t_n}) \in V'\right).$$

Now, assume there exists some $(u_0, v_0) \in V$ such that $\partial_v g(u_0, v_0) \neq 0$. By continuity, we can then find some open neighborhood $(u_0, v_0) \in V_0 \subset V$ on which $\partial_v g$ is different from 0. But, by Lemma 2 it would then follow that $\tilde{y}_t = g(y_t, z_t)$ is non-stationary which results in a contradiction. Thus, we must have $\partial_v g(u, v) = 0$ for all $(u, v) \in V$. Then,

since $\partial_v g$ is analytic, by the Identity Theorem (see, e.g., Krantz and Parks [2002]), it follows that $\partial_v g = 0$ everywhere as desired.

Part (ii): First note that f is a piece-wise diffeomorphic homeomorphism (see Lemma 1). In particular, there exists a countable collection of closed sets $(U_i)_{i\in\mathcal{I}}$ covering $D_{(y,z)}$ such that ∂U_i has Lebesgue measure 0 and f is diffeomorphic on $\text{int}(U_i)$ for each $i \in \mathcal{I}$ (Definition 11). Now, defining $\mathcal{O} := \bigcup_{i\in\mathcal{I}} U_i$ we find that $\mathcal{O}|_{e(U)}$ is dense in $e(U)$. If there exists some $(u_0, v_0) \in \mathcal{O}$ such that $\partial_v g(y_0, v_0) \neq 0$, then, by continuity, there is some open neighborhood $(u_0, v_0) \in V_0 \subset U_i$ such that $\partial_v g$ is different from 0 on V_0. But then, since the law of z_t is strictly non-stationary also when restricted to $\pi_z(V_0)$, Lemma 2 leads to a contradiction again. We therefore conclude that $\partial_v g(u, v) = 0$ for all $(u, v) \in \mathcal{O}$. □

While being small enough to yield identifiability for any strictly non-stationary and regular process, z_t, the class of analytic admissible mixings is still a significant generalization of the purely linear case. Indeed, modulo invertibility constraints, it includes affine functions, polynomials, neural networks with analytic activation functions, and many other types of functions. If one wants access to more flexible functions such as, for example, splines or neural networks with only piece-wise differentiable activation functions (the prime example being, of course, piece-wise linear functions such as the leaky ReLU activation function), then the class of analytic mixings no longer suffices. In other words, in order to guarantee identifiability, we need to put stronger assumptions on z_t. A good example to keep in mind of what we here refer to as *globally* non-stationary processes are non-stationary Gaussian processes such as Brownian motion.

3.3 Discerning stationarity using signatures

Having converged on a well-behaved definition of nonlinear cointegration, the natural next step is asking how the stationary embedding might be estimated given an observation of (possibly sub-sampled) trajectories of x. Naturally, the best we can hope for is being able to estimate $[e]_{\mathcal{F}}$, and (3.2.1) gives a suggestion for how to do so. It suffices to find a de-mixing f such that $f_{1:k}(x)$ is stationary since (3.2.1) then guarantees that $f \in [e]_{\mathcal{F}}$. Our approach, then, is similar to that of Schell and Oberhauser [2023] in that we seek to define some statistic discriminating stationary from non-stationary processes. In theory, one should be able to find a candidate for f by minimizing this objective over the given class of admissible mixings. This approach is different from another strand of BSS where one seeks to estimate the latent processes using contrastive learning [Hyvarinen and Morioka, 2016, 2017].

What we are looking for, is a well-behaved map $\varphi : \mathcal{M}_1(C(\mathbb{R}^p)) \to \mathbb{R}_+$ such that $\varphi(\mu) = 0$ if, and only if, $\mu \in \mathcal{S}^p(\tau)$ (where $\tau > 0$). Assuming, then, that we are given such an *oracle* φ and that $(\mathcal{F}, z)$ is identifiable, by Lemma 1, we find that, for τ sufficiently large,

$$\arg\min_{f\in\mathcal{F}(U)} \varphi\left(f_{1:k}(x)\right) \subset [e]_{\mathcal{F}}. \tag{3.3.3}$$

This section is devoted to developing such an oracle. The basic idea is to start with some way of discriminating laws of stochastic processes using a *statistical divergence*. Given such a discriminator, call it Φ, we can then check if the law μ of a given stochastic process $x \in C(\mathbb{R}^p)$ agrees with that of $\mu \circ \theta_t$ for all $t > 0$. This is encapsulated in the following lemma.

Lemma 1. *With $\tau > 0$, let $K \subset C(\mathbb{R}^p)$ and $\Phi : \mathbb{R}_+^2 \times \mathcal{M}_1(K) \to \mathbb{R}_+$ such that, for all $\mu \in \mathcal{M}_1(K)$, the map $(s,t) \mapsto \Phi(s,t,\mu)$ is continuous and 0 if, and only if, $(x_{s+r})_{r\in[0,\tau]}$ and $(x_{t+r})_{r\in[0,\tau]}$ are equal in law where $x \sim \mu$. Then,*

$$\varphi(\mu) := \limsup_{T\to\infty} \int_0^T \int_0^T \Phi(s,t,\mu) ds dt \tag{3.3.4}$$

satisfies $\varphi(\mu) = 0$ if and only if $x \in \mathcal{S}^p(\tau)$.

Proof. We shall prove the "only if"-direction since the other direction follows by assumption. Assume that $x \notin \mathcal{S}^p(\tau)$. Then, there exist some $s_0, t_0 \in J_T$ such that $\Phi(s,t,\mu) > 0$. But, by continuity, this implies that we can take some open interval $I \ni s_0$ such that $I \ni s \mapsto \Phi(s,t_0,\mu)$ is strictly positive which, in turn, must mean that $\varphi(\mu) > 0$. □

In line with previous notation, we shall often replace the law of a stochastic process with the random variable itself, that is, we write $\Phi(s,t,x)$ (resp. $\varphi(x)$) instead of $\Phi(s,t,\mu)$ (resp. $\varphi(\mu)$) when this does not otherwise cause any confusion.

3.3.1 Signature kernel divergence

To build our discriminator, Φ, we shall use a feature map stemming from rough path theory known as the *signature* [Cass and Salvi, 2024]. Since signatures characterize the law of stochastic processes, they will serve as a great tool in this endeavor. We can use signature kernels to construct a metric for laws of path-valued random variables. This metric is the *signature kernel maximum mean discrepancy* (MMD) [Salvi et al., 2021c]. We will lay out in detail how this can be done below. Suffice it to say for now that one is free to substitute the signature kernel MMD with any suitable divergence.

For ease of exposition, we now restrict our attention to $C_{1-var}(\mathbb{R}^p) \subset C(\mathbb{R}^p)$, that is, the space of paths of bounded variation[5]. On this space, we may define the signature $S : C_{1-var} \to C(T((\mathbb{R}^p)))$ which takes as input a bounded variation path and returns a continuous path taking values in the extended tensor algebra $T((\mathbb{R}^p)) = \prod_n (\mathbb{R}^p)^{\otimes n}$ (see also Appendix 3.C). For $x \in C_{1-var}(J_T, \mathbb{R}^d)$, the elements of $S(x)$ are given by the iterated tensor integrals of x_t with respect to itself. To be precise, $S(x) = (1, S_1(x), S_2(x), \dots)$ where

$$S_k(x) = \int \cdots \int_{0<u_1<\cdots<u_k<T} dx_{u_1} \otimes \cdots \otimes dx_{u_k}.$$

The signature can be viewed as a feature map playing the role of monomials on path-valued data [Lyons, 2014]. It inherits many properties of real-valued monomials such as

[5] All of this can be extended to the space of geometric rough paths which includes, for example, solutions to SDEs as the ones discussed below.

universality and characteristicness [Chevyrev and Oberhauser, 2022]. In particular, it distinguishes paths up to *tree-like equivalence* and translation, i.e., for two paths starting at 0, $S(x) = S(y)$ if and only if $x \sim_t y$ with $\sim_t$ being the equivalence relation given by tree-like equivalence (see, e.g., Definition 1.3 in Hambly and Lyons [2010]). Equipping the extended tensor algebra with the natural inner product we can also consider the *signature kernel* $k(x, y) = \langle S(x), S(y) \rangle$ which is a kernel on path-space. This leads to the signature kernel MMD,

$$\mathrm{MMD}_{sig}(x, y)^2 = \|\mathbb{E}S(x) - \mathbb{E}S(y)\|^2 = \mathbb{E}k(x, x') + \mathbb{E}k(y, y') - 2\mathbb{E}k(x, y)$$

where the pairs x, x' and y, y' are both i.i.d. We can then define φ_{sig} as in Eq. (3.3.4) with MMD_{sig} our choice of divergence. However, to ensure that MMD_{sig} is actually a metric, we make two minor modifications such that it is sensitive to translation by both tree-like paths and constants. In particular, for a path $x \in C_{1-var}([s, t], \mathbb{R}^p)$, we let $\tilde{x} \in C_{1-var}([s, t], \mathbb{R}^{p+1})$ be the path obtained by applying the *ι-augmentation* for $\iota : [s, t] \to [0, 1]$ the linear path connecting 0 and 1, i.e., $\tilde{x} = (x, \iota)$ (see also Definition 14). Furthermore we replace the signature in the definition of MMD_{sig} with the I-augmented signature, i.e., $S^I(x) = \exp(x_s) \otimes S(x)$ (see also Definition 15). We then let $\Phi_{sig}(s, t, x)^2 = \mathrm{MMD}_{sig}(\tilde{x} \circ \theta_s, \tilde{x} \circ \theta_t)^2$. Theorem 5 ensures that $(x, y) \mapsto \mathrm{MMD}_{sig}(\tilde{x}, \tilde{y})$ is a metric on any compact subset of $C_{1-var}(J_\tau, \mathbb{R}^p)$ so that Φ_{sig} satisfies all the requirements of Lemma 1. Combined with (3.2.1), we then obtain the following result reminiscent of Theorem 4 in Schell and Oberhauser [2023] for free.

Theorem 2. *Assume that the pair $(\mathcal{F}, z)$ is identifiable. Then, if x_t is $\mathcal{CI}_k$ with $e_2(x_t) = \pi_{(k+1):p}(z_t)$ and $(e_1, e_2) \in \mathcal{F}(U)$, we have that, for τ large enough,*

$$\arg\min_{g \in \mathcal{G}} \varphi_{sig}(g_{1:k}(x)) \subset [e]_{\mathcal{F}} \tag{3.3.5}$$

for any $\mathcal{G} \subset \mathcal{F}(U)$ such that $\mathcal{G} \cap [e]_{\mathcal{F}} \neq \emptyset$.

Remark 3. In many applications, depending on the nature of the latent non-stationary component, z_t, it may or may not be advantageous to consider different modifications of the signature. We applied the ι-augmentation and the I-augmentation above mainly to provide the necessary theoretical guarantees, that is, to obtain (3.3.5). Another often advantageous modification involves first lifting the path to some Hilbert space and then computing the signature kernel on the resulting Hilbert space valued paths. If one chooses this lift to be the feature map associated to a kernel function on Euclidean space, then different kernel tricks can be employed such that we may circumvent actually having to compute the lifted path. See, for example, Lee and Oberhauser [2023] or Appendix 3.C. In subsequent sections, we discuss both numerical considerations and perform empirical comparisons of different methods.

3.4 Estimating stationary embeddings

Rarely will we have access to the law of the stochastic process under consideration. Instead, one is often presented with one of two cases: 1) Multiple independent discrete

sub-samples or 2) a single long discretely sub-sampled trajectory. Noting that 2) is effectively a generalization of 1), we shall mainly focus on the latter case. Specifically, we observe the data $\mathbf{x}_n = (x_t)_{t\in\mathcal{T}_n}$ on a *time grid* $\mathcal{T}_n := \{0 = t_0^n < \cdots < t_n^n = T_n\}$. We shall make the following assumptions on the asymptotics of $\mathbf{x}_n$ as $n \to \infty$:

- The final time T_n increases, $T_n \to \infty$.
- The mesh of $\mathcal{T}_n$ goes to 0, $\|\mathcal{T}_n\| := \max_{k\leq n} |t_k^n - t_{k-1}^n| \to 0$.
- Each time grid includes the previous one, $\mathcal{T}_n \subset \mathcal{T}_{n+1}$.

In words: We let the length (in unit time) of the observation go to infinity and the fineness to 0. We note that not all three requirements are necessary for obtaining the results that follow below. Versions of the main results, Lemma 2 and Theorem 3, still hold with the according modifications. The focus of this section is finding an analog of $\varphi_{sig}(x)$ that works for streams such as $\mathbf{x}_n$. Towards this goal we define a *batching protocol*. The main idea is that signatures take as input entire *paths* (or sequences for the discrete signature). Thus, we need to consider the data at this level as well. We identify an observation with a slice of the observed sequence and to estimate the expected signature we then average a collection of such slices. Crucially, we need to make sure that the order of observations is preserved. The idea is very similar to the approach taken in Issa and Horvath [2023] where the authors use the signature kernel MMD for change point detection. See also Figure 3.4.1 for an illustration.

Definition 8 (Batching protocol)**.** Let $\mathcal{T} = \{0 = t_0 < \cdots < t_n = T\}$ for some $T > 0$ be a time grid. We make the following nested definitions:

(i) A *window* of size $k > 0$ of the time grid $\mathcal{T}$ is a subset $\mathbf{b} \subset \mathcal{T}$ consisting of k consecutive points in $\mathcal{T}$, i.e., $\mathbf{b} = \{t_j, \ldots, t_{j+k-1}\}$ for some $0 \leq j \leq n - k + 1$.

(ii) A *batch* of size $m > 0$ of the time grid $\mathcal{T}$ is a set of windows $\mathfrak{b} = \{\mathbf{b}_1, \ldots, \mathbf{b}_m\}$ such that, for each $2 \leq i \leq m$, $\min \mathbf{b}_{i-1} < \min \mathbf{b}_i$ and $\max \mathbf{b}_{i-1} < \max \mathbf{b}_i$.

(iii) Finally, a *batching protocol* of size $B > 0$ of the time grid $\mathcal{T}$ is a set of batches $\mathfrak{B} = \{\mathfrak{b}_1, \ldots, \mathfrak{b}_B\}$.

Now let $\mathfrak{B}_n = \{\mathfrak{b}_1^n, \ldots, \mathfrak{b}_{B_n}^n\}$ be a batching protocol of the time grid $\mathcal{T}_n$ where we write $\mathfrak{b}_k^n = \{\mathbf{b}_1^{n,k}, \ldots, \mathbf{b}_{B_{n,k}}^{n,k}\}$ and $\#\mathbf{b}_j^{n,k} = B_{n,k,j}$, i.e., B_n denotes the size of the batching protocol, $B_{n,k}$ the size of the kth batch, and $B_{n,k,j}$ the size of the jth window in the kth batch. Note that we allow the batching protocol to depend on the sample size n. This is crucial for deriving the asymptotic results presented in the subsequent section. For ease of notation, given a window $\mathbf{b}$, we write $x_{\mathbf{b}} = (x_t)_{t\in\mathbf{b}}$ and $x_{[\mathbf{b}]} := x_{[\min \mathbf{b}, \max \mathbf{b}]}$. For any time series $\mathbf{x} = (x_t)_{t\in\mathcal{T}}$ or stochastic process $(x_t)_{t\in J_T}$ and an appropriate function g, the notation $g(\mathbf{x})$ or $g(x)$ is to be understood as the time series or process resulting from a point-wise application of g. Furthermore, $S_{\iota,g}^I(x)$ denotes the composition $x \mapsto g(x) \mapsto \tilde{x}_g \mapsto S^I(\tilde{x}_g)$ where $\tilde{x}_g$ is the ι-augmentation of $x_g := g(x)$ for ι the simple linear path connecting 0 and 1 as in the previous section. We use a similar notation when $S(x)$ is

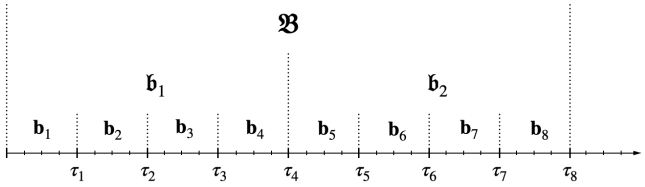

Figure 3.4.1: A batching protocol with two batches, each consisting of four windows. The $i'th$ window contains the observations between τ_{i-1} and τ_i. The batches in this illustration contain windows from two disjoint intervals, but this need not be the case in general.

replaced with the discrete signature (see Appendix 3.C.2) applied to $x_{\mathbf{b}}$, i.e., $\hat{S}^I_{\iota,g}(x_{\mathbf{b}})$.[6] We define the projection $\pi^{(m)} : T((\mathbb{R}^p)) \to \bigoplus_{k\leq m}(\mathbb{R}^p)^{\otimes k}$ onto the first m levels of the tensor algebra.

1. We only observe a discretely sub-sampled trajectory of x_t and, consequently, can not compute the exact signatures $S^I_{\iota,g}(x_{[t,t+\tau]})$. We shall approximate it with the discrete signature truncated of some order $m \geq 1$. Specifically, for each window $\mathbf{b}^{n,k}_j$,
$$S^I_{\iota,g}(x_{[t,t+\tau]}) \approx \pi^{(m)} \circ \hat{S}^I_{\iota,g}(x_{\mathbf{b}^{n,k}_j}) := \hat{\mathbf{s}}_m(g(\mathbf{x}_n); \mathbf{b}^{n,k}_j),$$

2. Next, since we do not have access to the distribution from which $\mathbf{x}_n$ is sampled, we are not able to compute $\mathbb{E}\hat{\mathbf{s}}(g(\mathbf{x}_n); \mathbf{b}^{n,k}_j)$. An obvious choice is then to estimate the mean by a sample average over a batch. In particular, for each batch $\mathfrak{b}^n_k$, we make the further approximations
$$\mathbb{E}S^I_{\iota,g}(x) \approx \frac{1}{B_{n,k}} \sum_{\mathbf{b}\in\mathfrak{b}^n_k} \hat{\mathbf{s}}_m(g(\mathbf{x}_n); \mathbf{b}) := \hat{\mathfrak{s}}_m(g(\mathbf{x}_n); \mathfrak{b}^n_k),$$

3. For each combination of a pair of two batches, $\mathfrak{b}^n_k$ and $\mathfrak{b}^n_l$, starting at s^n_k and s^n_l respectively, this then yields an approximation
$$\Phi_{sig}\left(s^n_l, s^n_k, g(x)\right) \approx \|\hat{\mathfrak{s}}_m(g(\mathbf{x}_n); \mathfrak{b}^n_k) - \hat{\mathfrak{s}}_m(g(\mathbf{x}_n); \mathfrak{b}^n_l)\|^2 := \hat{\Phi}_{sig}(g(\mathbf{x}_n); \mathfrak{b}^n_k, \mathfrak{b}^n_l).$$

4. As the last step, we can then obtain a final approximation by aggregating over all combinations of batches in the protocol, that is,
$$\varphi_{sig}(g(x)) \approx B_n^{-2} \sum_{\mathfrak{b},\mathfrak{b}'\in\mathfrak{B}_n} \hat{\Phi}_{sig}(g(\mathbf{x}_n); \mathfrak{b}, \mathfrak{b}') := \hat{\varphi}_{sig}(g(\mathbf{x}_n); \mathfrak{B}_n).$$

It might not be immediately obvious how $\hat{\varphi}_{sig}$ is related to φ_{sig}. Let us consider a specific batching collection. The arguably simplest batching protocol is the one given in

[6] Note that, with $z = g(x)$ and $\mathbf{z} = g(x_{\mathbf{b}})$, Remark 9 ensures that $\tilde{z}_{\mathbf{b}} = \tilde{\mathbf{z}}$.

Fig. 3.4.1. We first split up the sequence $\mathbf{x}_n$ into B_n windows of equal size n/B_n. We then define two batches containing the first and last half of the windows respectively. Our statistic then simply compares the average signature over the first batch and the second batch, i.e., we are effectively testing if the distribution is the same on both halves of the observed time frame. This is obviously different from checking whether the distribution is the same at *all times*. For example, consider what would happen if the distribution of x_t changes back and forth each window. Then, this specific batching protocol would not be able to detect the non-stationarity. Other, more complex batching protocols could, though. In general, choosing a batching protocol depends on the application at hand and can be seen as a sort of hyperparameter. Smaller window sizes mean that we can fit more windows into a batch of fixed length, but this comes at the cost of not being able to detect more complex long term changes in the distribution. Similarly, reducing the number of windows in a batch means that we can fit more batches into a batching protocol, but it will also lead to a higher variance in estimating the expected signature and, therefore, the maximum mean discrepancy (steps 2 and 3 above).

What we will show in the following section is that the approximation converges to its expectation if x_t is stationary and satisfies certain mixing conditions. Of course, when x_t is stationary, the expectation of $\hat{\varphi}_{sig}(x_t)$ is 0 and therefore equal to the oracle $\varphi_{sig}(x_t)$. Recall that our goal is to use $\hat{\varphi}_{sig}$ as a way to discriminate non-stationary processes. In particular, it will be enough to require that $\hat{\varphi}_{sig}$ remains strictly positive in the non-stationary case. This is harder to prove. We shall show that under specific assumptions on the kind of non-stationarity, this will hold for every $g(x)$ over a suitable class of transformations.

Remark 4. In some cases it might be a better idea not to average over all pairs of batches in our batching protocol. Actually, there is nothing in our formulation stopping us from only taking the average over a subset of pairs. Even more generally, one may consider an aggregator, $A_n : \mathbb{R}_+^{B_n \times B_n} \to \mathbb{R}_+$, and define

$$\hat{\varphi}_{sig}(g(\mathbf{x}_n); \mathfrak{B}_n, A_n) := A_n \left(\hat{\Phi}_{sig}(g(\mathbf{x}_n); \mathfrak{b}, \mathfrak{b}')_{\mathfrak{b},\mathfrak{b}' \in \mathfrak{B}_n} \right).$$

3.4.1 Consistency

This sections is devoted to studying the limiting behaviour of $\hat{\varphi}_{sig}$ under mixing conditions with and without stationarity. The main result concerns the average discrete signatures $\hat{\mathfrak{s}}_m(g(\mathbf{x}_n); \mathfrak{b})$ over a batch $\mathfrak{b}$ as estimators of $\mathbb{E}S^I_{\iota,g}(x)$. Most of the theory and all of the proofs have been relegated to Appendix 3.B. Left here are the most important results along with the appropriate underlying assumptions. We note that general asymptotic results and concentration inequalities exist for the MMD even for dependent data [Dehling and Wendler, 2010, Chérief-Abdellatif and Alquier, 2022], but in order to study how $\hat{\varphi}_{sig}$ behaves under the alternative of non-stationary data, Lemma 2 even will be quite useful as it only requires a mixing condition and not stationarity.

As mentioned above, the consistency results that we introduce all presuppose some sort of mixing conditions. These are universal in time series analysis since they allow us to transfer many of the classical asymptotic results to dependent data. For a thorough

introduction to this concept, we refer to Doukhan [2012]. For the sake of simplicity, we here only introduce the notion of α-mixing.

Definition 9 (α-mixing)**.** Let $X = (X_1, X_2, \dots)$ be a sequence of random variables (or time-series) and define the σ-algebras $\mathbb{X}_s^t := \sigma(X_u : s \leq u \leq t)$. The *$\alpha$-mixing coefficient* of X for $r \geq 1$ is defined as

$$\alpha_r(X) := \sup_{j \in \mathbb{N}} \sup_{\substack{A \in \mathbb{X}_1^j \\ B \in \mathbb{X}_{j+r}^\infty}} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|.$$

We call a time-series array $(X_t^n)_{n,t \geq 0}$ *α-mixing* or *strongly mixing* if the α-mixing coefficients go uniformly to zero, i.e., $\sup_n \alpha_r(X^n) \to 0$, as $r \to \infty$.

For a given batch, $\mathfrak{b}$, we define the average of the expected signatures

$$\mathfrak{s}(g(x); \mathfrak{b}) := \frac{1}{\#\mathfrak{b}} \sum_{\mathbf{b} \in \mathfrak{b}} \mathbb{E}\left(S_{\iota,g}^I(x_{[\mathbf{b}]})\right),$$

We can now state our result on the consistency of $\hat{\mathfrak{s}}_m(g(\mathbf{x}_n); \mathfrak{b}_n)$ as estimators of $\mathfrak{s}(g(x); \mathfrak{b}_n)$ where $\mathfrak{b}_n$ is some batch that also depends on the sample size. Since we are measuring the stationarity of transformations of $\mathbf{x}_n$ (recall (3.3.3)), we prove that consistency holds uniformly over a suitable class of nice transformations.

Lemma 2 (Signature consistency)**.** *Let $\mathcal{G} \subset C(\mathbb{R}^p, \mathbb{R}^q)$ compact and $\mathbf{x}_n = (x_t)_{t \in \mathcal{T}_n}$ with the time grids $\mathcal{T}_n$ as described in the beginning of the section. Assume that x, $\mathcal{G}$ and $\tau > 0$ satisfy the technical Assumption 2. Assume that $\mathbf{x}_n$ is strongly mixing. Let $\mathfrak{b}_n$ be a sequence batches of size B_n with $B_n \to \infty$ as n increases and such that each window is of length $\tau > 0$. Then, for any $\epsilon > 0$, there exists $m_0 \geq 1$ such that, for all $m \geq m_0$,*

$$\lim_{n \to \infty} \mathbb{P}\left(\sup_{g \in \mathcal{G}} \|\hat{\mathfrak{s}}_m(g(\mathbf{x}_n); \mathfrak{b}_n) - \mathfrak{s}(g(x); \mathfrak{b}_n)\| > \epsilon\right) \to 0.$$

Remark 5. Here we imposed a strong mixing assumption. This assumption is exclusively used to show that the discrete signature is in a sense *ergodic*, i.e., that taking the time-average corresponds asymptotically to taking the spatial average (see also Lemma 5). One may be able to prove this without the mixing assumption, but under some alternative condition. A particular avenue worthy of exploration that comes to mind is to instead require that x is a (bijective transformation of a) recurrent Markov chain as was done in, e.g., Karlsen et al. [2007]. We leave such efforts open for future work.

Recall that we are interested in solving (3.3.3) for some identifiable pair $(\mathcal{F}, z)$ with our objective function being φ_{sig}. Not having access to φ_{sig} we must resort to using the alternative objective $\hat{\varphi}_{sig}(\cdot; \mathfrak{B}_n)$ for an appropriate batching protocol $\mathfrak{B}_n$. Furthermore, in practice we usually will not perform the minimization over the entire set $\mathcal{F}(U)$, but only over some sufficiently flexible subset $\mathcal{G} \subset \mathcal{F}(U)$. Thus, our estimate of the stationary embedding, call it $\hat{e}$, solves

$$\hat{e} \in \arg\min_{g \in \mathcal{G}} \hat{\varphi}_{sig}(g_{1:k}(\mathbf{x}_n)) \tag{3.4.6}$$

Now, proving consistency of this estimator would then amount to proving that $\mathrm{d}_\infty(\hat{e}, [e]_\mathcal{F})$ goes to 0 in probability as the sample size increases. Here d_∞ denotes the uniform distance between two continuous functions and $\mathrm{d}_\infty(g, A) := \inf_{f \in A} \mathrm{d}_\infty(g, f)$ for any set of continuous functions A. The consequence of Lemma 2 is not quite enough to ensure that $\hat{\varphi}_{sig}$ stays asymptotically positive in probability for non-stationary processes. Indeed, once could imagine certain adversarial examples where $\|\mathfrak{s}(x; \mathfrak{b}) - \mathfrak{s}(x; \mathfrak{b}')\| = 0$ for all batches $\mathfrak{b}, \mathfrak{b}' \in \mathfrak{B}_n$ even if x is non-stationary. To rule out such cases we impose an additional assumption.

Assumption 1. *There exits some $\tau > 0$ and sequences of batches $\mathfrak{b}_n$ and $\mathfrak{b}'_n$ of size $B_n \to \infty$ such that each window is of length τ and, for every $\epsilon > 0$, there is some $c_\epsilon > 0$ such that, for all $g \in \mathcal{G}$ with $d_\infty(g, [e]_\mathcal{F}) \geq \epsilon$,*

$$\lim_{n \to \infty} \mathbb{P}\left(\hat{\Phi}_{sig}(g_{1:k}(\mathbf{x}_n); \mathfrak{b}_n, \mathfrak{b}'_n) > c_\epsilon\right) = 1.$$

This is a *very* high level assumption. We give some specific conditions under which Assumption 1 holds in Appendix 3.B.6. Notably, Assumption 1 holds if z_t is piece-wise stationary (see also Example 3.6.1). This, then, includes a lot of interesting settings encountered in practice such as in, for example, change-point detection [Truong et al., 2020] or independent component analysis [Hyvarinen and Morioka, 2016, Hyvarinen et al., 2019]. However, as can be seen from our experiments in Section 3.5, the statistic $\hat{\varphi}_{sig}$ works well empirically for a number of applications going beyond the asymptotic setting of piece-wise stationarity. We also note that in settings where we are given i.i.d. sample trajectories of x_t, we can make due without Assumption 1 since then we are able to approximating $\varphi_{sig}(x)$ is easy.

For $\mathcal{G} \subset C(\mathbb{R}^p)$ we define the set $[e_1]_\mathcal{G} := \pi_{1:k}([e]_\mathcal{F} \cap \mathcal{G})$.

Theorem 3. *Let x_t be $\mathcal{CI}_k$ with $x_t = d(y_t, z_t)$ and $\mathbf{y}_n = (y_t)_{t \in \mathcal{T}_n}$ strongly mixing with the time grids $\mathcal{T}_n$ as described in the beginning of Section 3.4.1. For $\mathcal{F}$ a class of admissible mixings with $d \in \mathcal{F}$ and $\mathfrak{B}_n$ a batching protocol, assume that $(\mathcal{F}, z)$ is identifiable and that x_t satisfies Assumption 1 for $\tau > 0$ and every $\mathfrak{b}, \mathfrak{b}'_n \in \mathfrak{B}_n$. Finally, for $\mathcal{G} \subset C(\mathbb{R}^p)$ such that every $g \in \mathcal{G}$ is of full rank, assume that $y, [e_1]_\mathcal{G}$, and τ satisfy the technical Assumption 2. Then, for any $\epsilon > 0$, there is a sufficiently large truncation level $m \geq 1$ such that, for $\hat{e}$ as in (3.4.6), we have*

$$\lim_{n \to \infty} \mathbb{P}\left(\mathrm{d}_\infty\left(\hat{e}, [e]_\mathcal{F}\right) > \epsilon\right) = 0.$$

Remark 6 (Numerics). The minimization in (3.4.6) will be solved numerically using some variant of stochastic gradient descent. It therefore requires computing $\hat{\varphi}_{sig}$ at each step of the optimizer. Crucially, then, the computational complexity of our objective severely impacts the feasibility of the current approach. Luckily, there are many ways to simplify the computation of $\hat{\varphi}_{sig}$. In particular, we shall draw inspiration from the recent work in Toth et al. [2023] and approximate the signature kernel with *random fourier signature features*. Two specific algorithms depending on the choice of batching protocol are given in Appendix 3.D.

3.5 Applications

In this section we go through some applications of stationary embeddings and consider a few special cases and extensions. This is by no means meant to be an exhaustive exploration, but only to serve as a key showcase of the kind of problems that can be tackled with this framework. First we consider the use of stationary embeddings for *dimensionality reduction* (or *manifold learning*) of stochastic processes. As we shall discuss further down below, the stationary embedding can be likened to invariant manifolds of deterministic dynamical systems. A special case is when the decoding is of the form $d(y,z) = d^s(y) + d^n(z)$ in which case a natural solution involves auto-encoders. Next, we consider how stationary embeddings can be used in regression problems where the predictor is non-stationary. We distinguish between stationary and non-stationary target variables. The former case lends itself well to a semi-supervised learning formulation.

3.5.1 Manifold learning

A canonical way to reduce the dimension of dynamical systems is via *invariant manifolds*. For an ordinary differential equation of the form $dx_t = f(x_t)dt$, an invariant manifold is a manifold $\mathbb{M}$ such that $x_s \in \mathbb{M}$ implies that $x_t \in \mathbb{M}$ for all $t \geq s$. Data driven methods seek to learn invariant manifolds from data and thereby significantly reduce the complexity of the system [Cenedese et al., 2022]. As the following result makes clear, nonlinear cointegration can serve as a natural generalization of invariant manifolds to SDEs or, more generally, to any stochastic process.[7] We note that invariant manifolds are also defined for random dynamical systems (see, e.g., Arnold [2013]) and the current view should therefore be considered as complementary to the existing dynamical systems litterature.

Proposition 1. *Let $x \in C(\mathbb{R}^p)$ be $\mathcal{CI}_k$ with $x_t = d(y_t, z_t)$ and d diffeomorphic on some open superset $U \supset D_{(y,z)}$. Define, for each $u \in D_y$, the open set U_u containing all $v \in \pi_{(k+1):p}(U)$ such that $(u,v) \in U$ along with $\mathbb{M}_u = d(u, U_u)$. Then, for each $u \in D_y$, the set $\mathbb{M}_u$ is a k-dimensional manifold embedded in $\mathbb{R}^p$ and the map $v \mapsto d(u,v)$ is an embedding.*

If we further let $a := \mathbb{E}(y_0)$ and assume that there exist $C, \alpha > 0$ such that, for each $v \in D_z$, the map $u \mapsto d(u,v)$ is α-Hölder continuous at $u = a$ with constant bounded by C, then, for any $\epsilon > 0$ and $t \geq 0$,

$$\mathbb{P}\left(\inf_{x' \in \mathbb{M}_a} \|x_t - x'\| > \epsilon\right) \leq \left(\frac{C}{\epsilon}\right)^{1/\alpha} \mathrm{Var}(y_0). \tag{3.5.7}$$

Proof. Since U_u is open it is, in particular, a submanifold of $\mathbb{R}^k$. Then, since $v \mapsto d(u,v)$ is an embedding (d being a diffeomorphism), we find that $\mathbb{M}_u = d(u, U_u)$ is indeed a submanifold of $\mathbb{R}^p$. The second part follows from the following chain of inequalities along

[7] It also partly justifies our terminology in calling e (or $d = e^{-1}$) a stationary embedding.

with Markov's Inequality: With $x' = d(a, z_t)$,

$$\inf_{x' \in \mathbb{M}_a} \|x_t - x'\| \leq \|d(y_t, z_t) - d(a, z_t)\| \leq C\|y_t - a\|^\alpha.$$

□

The right-hand side of (3.5.7) does not depend on t. In particular, the equation implies that x_t, in probability, stays close to the manifold $\mathbb{M}_a$ embedded in $\mathbb{R}^p$. Intuitively, this means that the process x_t lives close to a lower-dimensional manifold with deviations governed by the stationary process $y_t = e_1(x_t)$. In a sense, this is the best we can hope for when considering SDEs instead of ODEs since, supposing the diffusion term non-degenerate, there is always some chance that the driving noise pushes the state out of whatever manifold would have been invariant in the deterministic case. This is also similar to the view of linear cointegration as representing an equilibrium and the stationary component causing deviations from this equilibrium. Another important fact is that $(\mathcal{F}, z)$ being identifiable is equivalent to saying that the manifold $\mathbb{M}_a$ is identifiable up to diffeomorphisms. Of course, this is also in direct analogy to the linear case where the *non-stationary* subspace is identifiable.

Example 3.5.1 (Stereographic projection)**.** As our first example we consider the case where $x_t \in \mathbb{R}^3$ is given by $x_t = d(y_t, z_t)$ with (y_t, z_t) satisfying

$$dy_t = -y_t dt + \sigma_y dw_t, \quad dz_t = -\theta(t, z_t)dt + \sigma_z dw_t, \tag{3.5.8}$$

and $\theta : \mathbb{R}_+ \times \mathbb{R}^2 \to \mathbb{R}^2$, $\sigma = (\sigma_y, \sigma_z) \in \mathbb{R}^3$, and $w_t \in \mathbb{R}^{3\times 3}$ a standard Brownian motion. We shall consider the case where $\theta(t, z) = -(z_{1,t}, z_{2,t})$ for $t \leq T/2$ and $\theta(t, z) = 4\theta(0, z)$ for $t > T/2$. In particular, y_t is a stationary OU-process and z_t is a non-stationary process that is stationary on each of the intervals $[0, T/2]$ and $(T/2, T]$. We will keep the mixing transformation fixed such that $d(y, z_1, z_2) = (r - y)P^{-1}(z_1, z_2)$. where $r > 0$ is some large number and P^{-1} is the inverse of the stereographic projection, i.e.,

$$P^{-1}(z_1, z_2) = \left(\frac{2z_1}{1 + z_1^2 + z_2^2}, \frac{2z_2}{1 + z_1^2 + z_2^2}, \frac{z_1^2 + z_2^2 - 1}{z_1^2 + z_2^2 + 1} \right)$$

In other words, we shall assume that x_t stays close to the 2-dimensional sphere $r\mathbb{S}^2$ of radius r embedded in $\mathbb{R}^3$. Finally, we note that a stationary embedding is known to be $x \mapsto \|x\|^2$, i.e., a simple polynomial. We will thus limit ourselves to the function class $\mathcal{G} = \{g : \mathbb{R}^3 \to \mathbb{R} \mid g \text{ is a second order polynomial}\}$.

Given a sample $\mathbf{x} = (x_t)_{t \in \mathcal{T}}$ with $\mathcal{T}$ a time-grid of 2048 evenly spaced points between $t_0 = 10$ and $T = 100$, we estimate e_1 as in (3.4.6), resulting in $\hat{e}_1$. We can then compare samples from the estimated stationary components $\hat{y} = \hat{e}_1(x)$ with the true y. Similarly, since $\mathbb{M}_a = \{u \in \mathbb{R}^p \mid e_1(u) - a = 0\}$, we obtain an estimate for the manifold by replacing e_1 with $\hat{e}_1$ and a with $\mathbb{E}(\hat{y})$. The results are reported in Fig. 3.5.2. Evidently, we are able to recover the underlying stationary component or, equivalently, the manifold $\mathbb{M}_a$. For a detailed discussion of this experiment we refer to Appendix 3.E.1.

♠

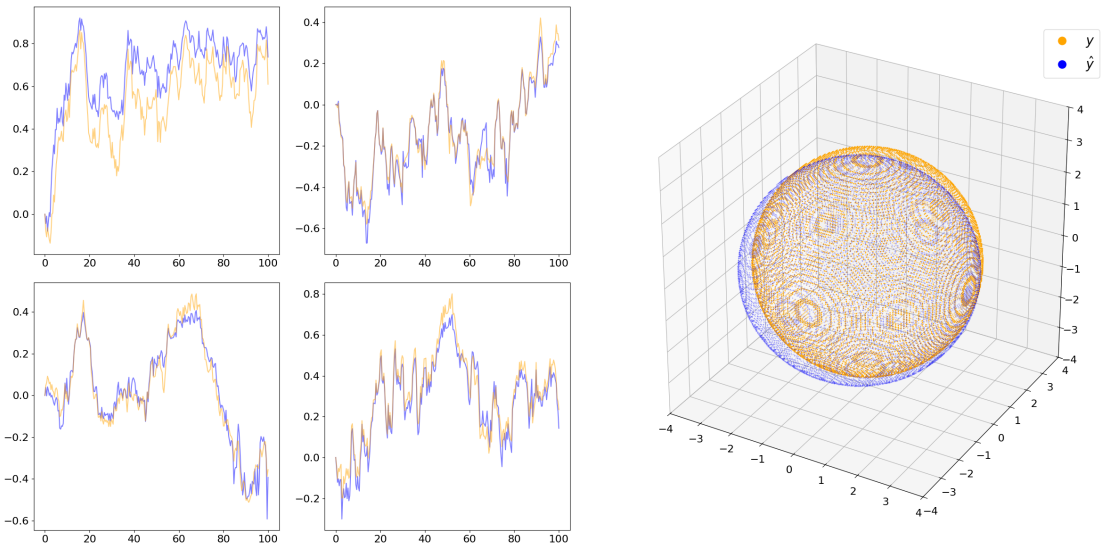

Figure 3.5.2: Estimating the stationary embedding of 3-dimensional non-stationary process. On the left are four samples of the true stationary component y_t (orange) along with the corresponding estimates $\hat{y}_t = \hat{e}_1(x_t)$ (blue). All lines have been translated to start at 0 and normalized to have range 1. On the right is the true lower-dimensional manifold $\mathbb{M} = r\mathbb{S}^2$ (orange) along with the estimate $\hat{\mathbb{M}} = \{x \in \mathbb{R}^3 \mid \hat{e}_1(x) - \sqrt{\hat{\mu}} = 0\}$ (blue) where $\hat{\mu}$ is the sample average of $\hat{y}$.

Example 3.5.2 (Additive mixings)**.** A model under which the Hölder continuity condition simplifies is the class of additive mixings, i.e., assuming that $d(y, z) = d^s(y) + d^n(z)$ where $d^s : \mathbb{R}^k \to \mathbb{R}^p$ and $d^n : \mathbb{R}^{p-k} \to \mathbb{R}^p$. In fact, we can then arrive at the following inequality even without Hölder continuity,

$$\mathbb{P}\left(\inf_{x' \in \mathbb{M}_a} \|x_t - x'\| > \epsilon\right) \leq \frac{\epsilon + C_K K}{\epsilon K} \mathrm{Var}(y_0) \tag{3.5.9}$$

where $K > 0$ such that $\|a\| \leq K$ and C_K is a constant such that d^s is C_K-Lipschitz on the ball around the origin of radius K.[8] Another nice property is that we can then write $d^s(y_t) = x_t - d^n(z_t) = x_t - d^n(e_2(x_t))$, that is, (3.3.3) amounts to solving the following auto-encoder problem

$$d^*, e^* \in \arg\min_{f,g} \varphi_{sig}(x - f(g(x))). \tag{3.5.10}$$

Consider now the same setup as in Example 3.5.1, that is, (y_t, z_t) are given by (3.5.8). We consider three specific choices of θ:

1. $\theta_1(t, z) = 0$ so that z_t is just a Brownian motion.
2. $\theta_2(t, z) = \sin(8t\pi/T)$ so that z_t is a Brownian motion plus a periodic drift.

[8] We note that such a constant exists by compactness and the fact that d^s is locally Lipschitz by Lemma 1.(i).

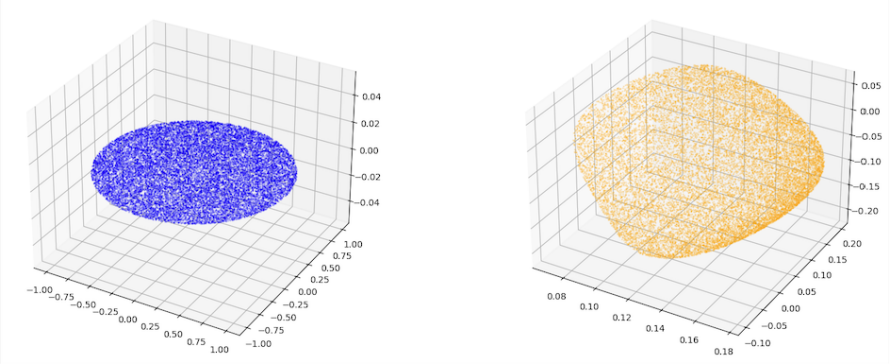

Figure 3.5.3: Left: The unit disk embedded in $\mathbb{R}^3$. Right: d^n applied to the unit disk.

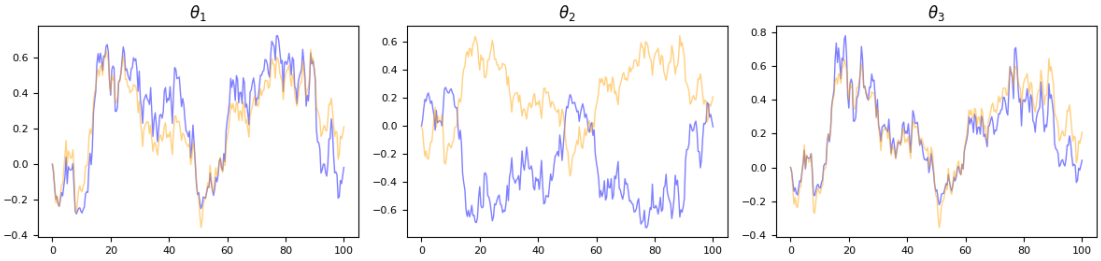

Figure 3.5.4: Estimating the stationary embedding of 3-dimensional non-stationary process. For each choice of θ, we plot a sample of the true stationary component y_t (orange) along with the corresponding estimate $\hat{y}_t = \hat{e}_1(x_t)$ (blue). All lines have been translated to start at 0 and normalized to have range 1.

3. $\theta_3(t, z) = -4z$ for $t < T/2$ and $\theta_3(t, z) = -z/4$ for $t \geq T/2$, i.e., z_t is a piece-wise Ornstein-Uhlenbeck process.

We assume that d^s and d^n are random initializations of multi-layer perceptrons (MLPs) with tanh activation functions and define $\mathbb{M} = d_2(0, \mathbb{R}^2)$. We refer to Figure 3.5.3 to see how the specific realization of d_2 maps the unit disk into $\mathbb{R}^3$.

Given an observation $\mathbf{x}$ as above, but of length 1024, we then estimate $\hat{e}_1 = \pi_1 \circ d^* \circ e^*$ where d^*, e^* solve (3.5.10). As seen in Figure 3.5.4, we are also in this case able to identify the stationary embedding (up to monotone transformations). We note that, for θ_2, flipping the sign of the $\hat{y}_t$ would cause the two lines to overlap as in the other two plots. This experiment was repeated for two other realizations of d^s and d^n with similar results. For a more detailed explanation of this experiment along with extra results we refer to Appendix 3.E.2.

♠

3.5.2 Regression with a non-stationary regressor

Another application is the problem of regressing some target process $v_t \in \mathbb{R}^q$ on a non-stationary process $x_t \in \mathbb{R}^p$ both of which are observed. One may consider two different

scenarios corresponding to whether the target process is stationary or not. We shall treat each of these separately.

Example 3.5.3 (Spurious regression)**.** In the case where v_t is non-stationary one has to be very careful in regressing v_t on x_t since this might result in a spurious relationship [Lee et al., 2005, Phillips, 2009, Tu and Wang, 2022], i.e., we might conclude that there is a strong connection between the two processes even though they are not related at all. We can posit the non-linear cointegration model $v_t = f(x_t) + u_t$ with u_t stationary in which case a regression makes sense. To ensure that we avoid spurious relationships we should then check whether this model is plausible. One naive way to do so would be to check how large the minimum in (3.4.6) is over some suitably flexible function class $\mathcal{G}$. The closer to 0 the objective is, the more evidence there is for the nonlinear cointegration model. ♠

Example 3.5.4 (Stationary target)**.** For simplicity, we let $p = q = 12$. Assuming that v_t is stationary, we then find that $f(x_t) = v_t - u_t$ is stationary as the difference of two stationary processes. Given sufficient regularity of f, it follows that x_t is not strictly non-stationary (see Definition 4) or, in other words, that x_t is $\mathcal{CI}_k$ of some order $k < p$ with $x_t = d(y_t, z_t)$. We may then write $v_t = g(y_t) + u_t$ for some suitable $g : \mathbb{R}^k \to \mathbb{R}^p$. Intuitively, this suggests that, in order to regress v_t on x_t, it might be beneficial to first learn the stationary embedding e and then regress v_t on $y_t = e_1(x_t)$. This is especially true if we only have access to a few observations of v_t, but more observations of x_t. We might then look at it as a problem of *semi-supervised* learning.

Consider then the following example: x_t is $\mathcal{CI}_k$ with $x_t = d(y_t, z_t)$ and $v_t = g(y_t) + u_k$. We consider different choices of k, but assume that it is known throughout. We assume that y_t is a stationary OU-process and that z_t is either a Brownian motion (BM), a Brownian motion plus a periodic drift (BM + P), or a stationary OU-process plus a periodic drift (OU + P). We let d be a randomly initialized MLP with tanh activation functions and $g(y) = \tanh(Ay+b)$ for some suitable $A \in \mathbb{R}^{p\times k}$ and $b \in \mathbb{R}^p$. Assuming now that we have observations $\mathbf{x} = (x_t)_{t\in\mathcal{T}}$ and $\mathbf{v} = (v_t)_{t\in\mathcal{T}_l}$ where $\mathcal{T} = \{t_0 < \cdots < t_n = T\}$ and $\mathcal{T}_l \subset \mathcal{T}$ is given by the first l time points in $\mathcal{T}$, we seek to learn the map $x \mapsto g(e_1(x))$ in three different ways. The first way splits the problem in two. One uses all of $\mathbf{x}$ to learn the stationary embedding (or the encoder) $\hat{e}_1 : x \mapsto \hat{y}$. We can then learn the decoder $\hat{g} : \hat{y} \mapsto \hat{v}$ by using the learned stationary component $\hat{y}_t$ as a proxy for the true latent y_t. This corresponds to the `STEMd` column in Table 3.5.1. The second method directly learns an autoencoder, but penalizing the encoding over all of $\mathbf{x}$ with $\hat{\varphi}_{sig}$ ensuring that the learned encoding behaves like a stationary process. We call this approach `STEMae` in Table 3.5.1. Finally, we learn a simple autoencoder by regressing $\mathbf{z}$ on the first l time points of $\mathbf{x}$. This last method does not use any of the information contained in $\mathcal{T} \setminus \mathcal{T}_l$. We will denote it by `autoencoder` in Table 3.5.1. We also include a model that learns the decoder directly on the true latent process y_t as a point of reference noting that its performance is unrealistic since we do not have access to y_t. Throughout we let $n = 1024$ and $l = 64$, that is, l is much smaller than n. We report the mean squared error on the unlabeled data $\mathcal{T} \setminus \mathcal{T}_l$ for all three methods and three different values of the cointegration rank k in Table 3.5.1. We observe that `STEMae` outperforms the simple autoencoder in

all cases. Furthermore, in many cases we obtain significant gains by first using all of $\mathbf{x}$ for learning the stationary embedding indicating that the current framework can indeed be employed for semi-supervised learning when regressing a stationary process on a non-stationary process. For more details on the exact setup of this simulation experiment we refer to Appendix 3.E.3.

		`STEMd`	`STEMae`	`autoencoder`	Oracle
$k = 3$	BM	**.0335**	.0411	.0598	*.0174*
	BM + P	**.0333**	.0444	.0812	*.0174*
	OU + P	**.0586**	.0757	.0759	*.0174*
$k = 2$	BM	**.0291**	0355	.0596	*.0141*
	BM + P	.0359	**.0327**	.1253	*.0141*
	OU + P	.0608	**.0385**	.0845	*.0141*
$k = 1$	BM	**.0217**	.0222	.0309	*.0133*
	BM + P	**.0218**	.0228	.0280	*.0133*
	OU + P	.0254	**.0183**	.0497	*.0133*

Table 3.5.1: Mean squared test error in regressing v_t on x_t with and without learning of the stationary embedding across different cointegration ranks k. A lower score is better. For each row, the best-performing method is **bold**. The "Oracle" corresponds to knowing the true latent stationary component y_t and serves as an unachievable benchmark.

3.6 Discussion

We have established a sensible notion of non-linear cointegration by viewing it as a blind source separation problem. In particular, our definition generalizes cointegration relationships to include non-linear functions as well. We call such non-linear cointegration relationships stationary embeddings. We have shown that similar identifiability guarantees hold for stationary embeddings as hold in the linear setting. Thus, we have obtained a well-defined target of inference.

Furthermore, we have discussed how one might go about estimating this target. One way to do so is by minimizing a statistic that discriminates stationary from non-stationary processes. Using the signature transform, a concept from rough path theory, we have built such a statistic as well as an approximation that is consistent under the null of stationarity. Consistency is derived under the asymptotic regime corresponding to observing an increasingly long and fine discrete sub-sample.

Finally, we have given a range of applications and shown the usefulness of the present framework through simulation experiments. In particular, similar to how linear cointegration is an equilibrium relationship with stationary deviations, stationary embeddings describe a lower dimensional manifold on which the process is supported up to deviations given by the stationary component.

Many questions are still unresolved. For one, we have not at all considered how to choose the cointegration rank k. A possible way to do so would be relying on a test similar to the linear case. Thus, deriving a test of nonlinear cointegration would be of the utmost importance. It would seem possible to base such a test on the statistic $\hat{\varphi}_{sig}$, but this would require determining the asymptotic distribution under the null. Related to this, is the problem of verifying Assumption 1 under more diverse settings since this would prove that the test has power to a wider range of alternatives. Empirically, more exhaustive experiments are necessary and especially on non-simulated and high-dimensional data. A possible avenue worthy of exploration is video data.

Appendix

In this appendix we collect some of the more technical details of the paper. The first section contains all the additional details related to the definition of stationary embeddings and identifiability. In Section 3.B we lay out the proof of the consistency results in Lemma 2 and Theorem 3. Since these results rely on many of the properties of signatures and their discrete counterpart, we dedicate Section 3.C to these questions. The section also cotains a few new ideas related to discrete signatures. In Section 3.D we discuss the numerical implementation of the proposed methodology. In particular, we introduce the random Fourier signature features of Toth et al. [2023] and develop the two main algorithms. Finally, in Section 3.5 we provide additional information on the simulation experiments of Section 3.5.

3.A. Technical Details

We split this section into two parts. The first defines *function classes* and what it means for a function to have a property only *piece-wise*. The second part deals directly with some auxiliary results necessary for the proofs in Section 3.2 and 3.2.1. It contains some quite general results regarding transformation of multivariate non-stationary processes. Especially Conjecture 1 is of importance since it allows us to show that the class of strictly non-stationary processes is quite general.

3.A.1 Function classes

Definition 10 (Function class)**.** A function class $\mathcal{F}$ is a collection of pairs (U, f) where U is an open subset of $\mathbb{R}^p$ (where p may differ across elements) and f is a function on U of dimension p. We write $\mathcal{F}(U)$ to denote all functions f on U such that $(U, f) \in \mathcal{F}$. We say that $\mathcal{F}$ is of property $\mathcal{P}$ if for all $(U, f) \in \mathcal{F}$ the function f is of property $\mathcal{P}$ on U.

Important kinds of function class that we revisit throughout are those satisfying some property piece-wise. We now precisely define what it means for a function $f : \mathbb{R}^p \to \mathbb{R}^p$ to have some property $\mathcal{P}$ piece-wise.

Definition 11 (Piece-wise $\mathcal{P}$)**.** Let $V \subset \mathbb{R}^p$ open and path-connected. We say that a function $f : V \to \mathbb{R}^p$ is piece-wise $\mathcal{P}$ for some property $\mathcal{P}$, if there exists a countable collection of path-connected closed sets $(U_i)_{i\in\mathcal{I}}$ covering V such that the following holds:

(i) The property $\mathcal{P}$ holds on $\text{int}(U_i)$ for all $i \in \mathcal{I}$.

(ii) The boundary ∂U_i has Lebesgue measure 0 for all $i \in \mathcal{I}$.

(iii) If $K \subset V$ compact, then K only intersects U_i for finitely many $i \in \mathcal{I}$.

This definition might seem somewhat convoluted, but it covers many (if not all) the usual cases. For example, if $p = 1$ and f being piece-wise linear usually means that f

is linear on a collection of non-empty open intervals whose closure contains the entire domain. This falls under the framework of Def. 11. Note, in particular, that the last condition implies that, for any $x, y \in V$ and $\gamma : [0,1] \to V$ smooth connecting x and y, it holds that $\gamma([0,1])$ only intersects U_i for finitely many $i \in \mathcal{I}$.

Recall that a function class is a class of admissible mixings if it is closed under composition and inversion. The following lemma gives a general way to construct classes of admissible mixings starting with a property $\mathcal{P}$ that only needs to hold piece-wise. In particular, it shows that $\mathcal{D}$ is a class of admissible mixings.

Lemma 1. *Let $\mathcal{P}$ be a property of functions that is closed under composition and inversion. Let $\mathcal{F} \subset \mathcal{D}$ denote the class of functions that are piece-wise $\mathcal{P}$ and $U \subset \mathbb{R}^p$ open and path-connected. Then:*

(i) $\mathcal{F}(U)$ is a subset of the set of locally Lipschitz functions on U.

(ii) If $f \in \mathcal{F}(U)$, then $f^{-1} \in \mathcal{F}(f(U))$.

(iii) If $f \in \mathcal{F}(U)$ and $g \in \mathcal{F}(f(U))$, then $g \circ f \in \mathcal{F}(U)$.

Proof. Let $f \in \mathcal{F}(U)$. For (i) it suffices to show that f is Lipschitz on any compact convex $K \subset U$. Take $(U_i)_{i\in\mathcal{I}}$ as in Def. 11 and let $U_{i_1}, \dots, U_{i_n}$ be such that $U_{i_j} \cap K \neq \emptyset$ and $K \subset \bigcup_j U_{i_j}$. We can then find $L > 0$ such that $\|Df(x)\| \leq L$ for any $x \in \operatorname{int}(U_{i_j} \cap K)$. For $x, y \in K$ let $\gamma : [0,1] \to K$ be the straight line connecting x and y. Assume without loss of generality that $x \in U_{i_1}$ and define $t_1 := \min\{\gamma^{-1}(\partial U_{i_1} \cap \partial K) \cap (0,1]\}$ and $x_2 = \gamma(t_1) \in \partial U_{i_1}$. Now, iteratively define $t_l := \min\{\gamma^{-1}(\bigcup_j \partial U_{i_j} \cap \partial K) \cap (t_{l-1}, 1]\}$ and $x_{l+1} = \gamma(t_l)$ for $l \geq 2$. Since $\gamma([0,1])$ only intersects finitely many U_{i_j}, there is some $m \in \mathbb{N}$ such that $x_m = y$. The sequence $x =: x_1, \dots, x_m = y$ then lies on the straight line connecting x and y and satisfies $x_l, x_{l+1} \in U_{i_{j_l}}$ for some $j_l \leq n$ and all $l \leq m-1$. But then, by the Fundamental Theorem of Calculus,

$$\|f(x) - f(y)\| \leq L \sum_{1\leq l\leq m-1} \|x_{l+1} - x_l\| = L\|x - y\|$$

where the last equality follows from the fact that all x_l lie on the same straight line. This proves that f is Lipschitz on K as wanted.

For (ii), define the collection $(V_i)_{i\in\mathcal{I}}$ where $V_i := f(U_i)$. First note that as the image under a homeomorphism of a closed path-connected set, V_i is itself closed and path-connected. Since the property $\mathcal{P}$ is closed under inversion on $\operatorname{int}(U_i)$, we find that f^{-1} is also of property $\mathcal{P}$ on $\operatorname{int}(f(U_i))$ (note that f is open since it is a homeomorphism). Furthermore, since f is locally Lipschitz, it maps null-sets to null-sets. Thus, $\partial V_i = f(\partial(U_i))$ has Lebesgue measure 0 for all $i \in \mathcal{I}$. Finally, let $K \subset f(U)$ be compact. Then, also $f^{-1}(K)$ is compact and therefore only intersects finitely many U_i. It follows that K then also intersects V_i for only finitely many $i \in \mathcal{I}$. This proves that f^{-1} is piece-wise of property $\mathcal{P}$ and, hence, $f^{-1} \in \mathcal{F}(f(U))$.

For (iii) first note that compositions of continuous (resp. invertible) functions are continuous (resp. invertible). Thus, $g \circ f$ is continuous with continuous inverse. Now

let $U_i \subset U$ and $V_j \subset f(U)$ be as in Def. 11. For all $(i,j) \in \mathcal{I} \times \mathcal{J}$, we define $W_{i,j} = U_i \cap f^{-1}(V_j)$ and note that $W_{i,j}$ is closed and path-connected. Since $\text{int}(W_{i,j}) \subset \text{int}(U_i)$ and $f(\text{int}(W_{i,j})) \subset \text{int}(V_j)$ we see that $g \circ f$ also satisfies $\mathcal{P}$ on $W_{i,j}$. Now, using the fact that

$$\partial W_{i,j} \subset \left(\partial U_i \cap f^{-1}(V_j)\right) \cup \left(U_i \cap f^{-1}(\partial V_j)\right),$$

we see that the Lebesgue measure $\partial W_{i,j}$ is 0 if we can show that $f^{-1}(\partial V_j)$ is a Lebesgue null-set. But this follows from the (i) and (ii) combined with the fact that locally Lipschitz functions map null-sets into null-sets. Finally, let K be compact. Then, it only intersects finitely many U_i. Also, since $f(K \cap U_i)$ is compact for every $i \in \mathcal{I}$, it only intersects finitely many V_j or, in other words, $K \cap U_j$ only intersects finitely many $f^{-1}(V_j)$. Of course, this then implies that $K \cap W_{i,j} \neq \emptyset$ for finitely many $(i,j) \in \mathcal{I} \times \mathcal{J}$. This shows that $g \circ f$ is piece-wise $\mathcal{P}$ on U and, as a result, $g \circ f \in \mathcal{F}(U)$. □

Remark 7. For the proofs of part (ii) and (iii) above we neglected to show that the inverse (resp. composition) is also piece-wise diffeomorphic (as would be necessary for $\mathcal{F} \subset \mathcal{D}$). This, however follows readily upon a slight modification of the proof. Namely, we can just replace the property $\mathcal{P}$ with the property $\tilde{\mathcal{P}}$ of being both diffeomorphic and $\mathcal{P}$ (inverse and composition of diffeomorphic functions being diffeomorphic).

3.A.2 Auxiliary results

First, a criterion yielding a wide range of strictly non-stationary processes.

DISCLAIMER: Another lemma originally stood in the place of the following conjecture. Unfortunately, I discovered a mistake in the proof of that lemma and a counterexample to the claim. This discovery was made only a couple of days prior to the submission deadline and I have therefore not been able to give a proof of the new conjecture. In any case, this concerns only the existence of strictly non-stationary processes which, admittedly, is quite a crucial part of the definition of stationary embeddings.

Conjecture 1. *For any $p \geq 1$ there exists some $q \geq 1$ and random variables $X_1, \dots, X_q \in \mathbb{R}^p$ supported on $D_{x_1}, \dots, D_{x_q}$ such that, for any open $\mathcal{O} \supset \bigcup_i D_{x_i}$ and differentiable $f : \mathcal{O} \to \mathbb{R}$ of constant rank 1, it holds that $f(X_1) \neq_d f(X_i)$ for at least one $i = 2, \dots, q$.*

A useful fact is that any strictly non-stationary process remains strictly non-stationary when conditioning on a stationary process. This is what ensures that our definition of strict cointegration is minimal in terms of the cointegration rank k. Before stating our next result, a brief remark on questionable notation. In line with the convention we have used so far, for two random variables X and Y, we shall write $X|Y$ to indicate a regular conditional probability distribution of X conditioned on Y. If X and Y have a joint density, $\rho_{X,Y}$, wrt. Lebesgue measure, then $X|Y$ can be characterized by the density $\rho_{X|Y}(x|y) = \rho_{X,Y}(x,y)/\rho_Y(y)$ which is defined for P_Y-almost all y.

Lemma 2. *Let $z \in C(\mathbb{R}^p)$ be strictly non-stationary and $y \in C(\mathbb{R}^q)$ stationary with (y_t, z_t) regular. Then, for any open $\mathcal{O} \supset D_{(y,z)}$ and continuous piece-wise differentiable $g : \mathcal{O} \to \mathbb{R}$ such that $\partial_z g(u,v) \neq 0$ for almost all $(u,v) \in D_{(y,z)}$, it holds that $w_t := g(y_t, z_t)$ is non-stationary.*

Proof. Throughout, for some regular stochastic process $x \in C(\mathbb{R}^p)$ and $\mathbf{t} \in \Delta_m([0,\infty))$, we let $\rho_X^{\mathbf{t}}$ denote the density of the Euclidean random variable $x_{\mathbf{t}} = (x_{t_1}, \dots, x_{t_m})$.

Since (y_t, z_t) is regular, (w_t, y_t) admits a density wrt. Lebesgue measure. For any $\mathbf{t} \in \Delta_m([0,\infty))$, we may then define the conditional density

$$\rho_{W|Y}^{\mathbf{t}}(\mathbf{w}|\mathbf{y}) = \frac{\rho_{W,Y}^{\mathbf{t}}(\mathbf{w},\mathbf{y})}{\rho_Y^{\mathbf{t}}(\mathbf{y})}$$

and analogously, the conditional density $\rho_{Z|Y}^{\mathbf{t}}(\mathbf{z}|\mathbf{y})$. We note that $\rho_{W|Y}^{\mathbf{t}}(\cdot|\mathbf{y})$ is the transformation of $\rho_{Z|Y}^{\mathbf{t}}(\cdot|\mathbf{y})$ under the map $G_{\mathbf{y}} : \mathbf{z} \mapsto (g(y_1, z_1), \dots, g(y_m, z_m))$. But then, z_t being striclty non-stationary would imply that there exists some $\tau > 0$ such that

$$\rho_{W|Y}^{\mathbf{t}} \neq \rho_{W|Y}^{\mathbf{t}+\tau}.$$

This, in turn, by stationarity of y_t, implies that the joint density $\rho_{W,Y}^{\mathbf{t}} = \rho_{W|Y}^{\mathbf{t}} \rho_Y^{\mathbf{t}}$ is non-stationary which concludes the proof. □

3.B. Consistency

This section is devoted to establishing the results in Section 3.4.1. Much of the following relies rather directly on the results of Section 8 in Schell and Oberhauser [2023]. There are a few key differences, however. The main one is that we are not considering the signature of the linear interpolation of discretely sub-sampled path, but the *discrete signature* (see Appendix 3.C.2). Although our main reasons for doing so are numerical, it turns out that this simplifies the theory in some places, as well. Furthermore, their ergodicity results rely on a stationarity assumption which we can not make in the present setting for obvious reasons.

In establishing Theorem 3 the majority of our efforts will be spent on showing that the average discretized signatures approximate the average expected signatures, i.e., establishing Lemma 2. By a simple triangle inequality, we find that, for all $m \geq 1$, $g \in \mathcal{G}$, and batches $\mathfrak{b}_n$,

$$\begin{aligned}\|\mathfrak{s}(g(x);\mathfrak{b}_n) - \hat{\mathfrak{s}}(g(\mathbf{x}_n);\mathfrak{b}_n)\| &\leq \|\mathfrak{s}(g(x);\mathfrak{b}_n) - \mathfrak{s}_m(g(x);\mathfrak{b}_n)\| \\ &+ \|\mathfrak{s}_m(g(x);\mathfrak{b}_n) - \mathbb{E}\hat{\mathfrak{s}}_m(g(\mathbf{x}_n);\mathfrak{b}_n)\| + \|\mathbb{E}\hat{\mathfrak{s}}_m(g(\mathbf{x}_n);\mathfrak{b}_n) - \hat{\mathfrak{s}}_m(g(\mathbf{x}_n);\mathfrak{b}_n)\| \quad \text{(3.B.1)}\end{aligned}$$

Intuitively, we can split up the residual term in (3.B.1) into three components each of which will be treated separately below. The first term corresponds to the *truncation error* resulting from truncating the signature at some finite order $m \geq 1$. The second term is the *discretization error* of only observing a sub-sampled version of the true underlying continuous time process. Finally, the third term is the *finite sample error* resulting from the fact that we do not have access to the underlying data generating process, but only a single long trajectory.

Throughout we require an assumption of a more technical nature regarding the class of candidate stationary embeddings $\mathcal{G} \subset C(\mathbb{R}^p)$ and x similar to Assumption 2 in Schell and Oberhauser [2023].[9] For a function $g \in \mathcal{G}$, we define the map $S_{\iota,g}^I : C_{1-var}([s,t],\mathbb{R}^p) \to$

[9] See also their Remark E.1 for a discussion.

$T((\mathbb{R}^{q+1}))$ as the composition $x \mapsto g(x) \mapsto \tilde{x}_g \mapsto S^I(\tilde{x}_g)$ where $\tilde{x}_g$ denotes the ι-augmented version of $g(x)$ for $\iota(r) = (r-s)/(t-s)$. $\hat{S}^I_{\iota,g}$ is defined analogously for the discrete signature applied to some sub-sampled version $\mathbf{x} = x_{\mathbf{b}}$.

Assumption 2. *Let $x \in C_{1-var}(\mathbb{R}^p)$ and, for some open $\mathcal{O} \supset D_x$, $\mathcal{G} \subset C(\mathcal{O}, \mathbb{R}^p)$ equipped with the topology of compact convergence. Let $\tau > 0$. Then, we require that $g(x) \in C_{1-var}(\mathbb{R}^p)$ for all $g \in \mathcal{G}$ and*[10]

$$\lim_{\delta \to 0} \sup_{g \in \mathcal{G}} \sup_{t \geq 0} \|g(x)\|_{[t,t+\delta],1-var} = 0. \tag{3.B.2}$$

Furthermore, for all $m \geq 1$, we shall assume the following moment bound

$$\mathbb{E}\left(\sup_{g \in \mathcal{G}} \sup_{0 \leq t} \left\{ \|g(x)\|_{[t,t+\tau],1-var} + \left\| \pi^{(m)} \circ \exp(g(x_t)) \right\| \right\} \right) < \infty \tag{3.B.3}$$

On the level of the signature, we assume that, for all $g \in \mathcal{G}$ and $t \geq 0$, the expected signature $\mathbb{E}S^I_{\iota,g}(x_{[t,t+\tau]})$ exists with some $\lambda > 1$ such that $\sup_{g \in \mathcal{G}} \sup_{0 \leq t} \mathbb{E}\|S^I_{\iota,g}(x_{[t,t+\tau]})\|_\lambda < \infty$.

We note that, since the signature is Lipschitz continuous on sets of bounded 1-variation (see, e.g., Proposition 7.66 in Friz and Victoir [2010]), (3.B.3) in particular implies that, for all $m \geq 1$,

$$\mathbb{E}\left(\sup_{g \in \mathcal{G}} \sup_{0 \leq t} \left\| \pi^{(m)} \circ S^I_{\iota,g}(x_{[t,t+\tau]}) \right\| \right) < \infty. \tag{3.B.4}$$

3.B.1 Truncation error

Lemma 1. *Assume that x, $\mathcal{G}$ and $\tau > 0$ satisfy Assumption 2. Then, for any $\epsilon > 0$, there is some $m_0 \geq 1$ such that, for all $m \geq m_0$,*

$$\sup_{g \in \mathcal{G}} \sup_{0 \leq t} \left\| \mathbb{E}S^I_{\iota,g}(x_{[t,t+\tau]}) - \mathbb{E}\left(\pi^{(m)} \circ S^I_{\iota,g}(x_{[t,t+\tau]}) \right) \right\| < \epsilon.$$

Proof. By Assumption 2, there is some $M \geq 0$ such that $\|\mathbb{E}S^I_{\iota,g}(x_{[t,t+\tau]})\|_\lambda \leq M$ for all $g \in \mathcal{G}$ and $0 \leq t$. But then, with $B_M := \{v \in T((\mathbb{R}^q)) \mid \|v\|_\lambda \leq M\}$, by Lemma D.1.(vi) in Schell and Oberhauser [2023] and linearity of the projection operator, we find that

$$\sup_{g \in \mathcal{G}} \sup_{0 \leq t} \left\| \mathbb{E}S^I_{\iota,g}(x_{[t,t+\tau]}) - \mathbb{E}\left(\pi^{(m)} \circ S^I_{\iota,g}(x_{[t,t+\tau]}) \right) \right\| \leq \sup_{v \in B_M} \|v - \pi^{(m)}(v)\| \to 0$$

for $m \to \infty$. □

[10] This statement is to be interpreted as holding path-wise, i.e., for any realization of the path, $x(\omega)$.

3.B.2 Discretization error

Of the following two lemmas the first is little more than a straight-forward combination of Corollary 4.4 in Király and Oberhauser [2019] and Definition 14. It gives a useful bound on the difference between the continuous signature and the discrete signature of a sub-sampled path in terms of the maximum 1-variation between sub-sampling increments. With this bound we can quite easily prove the main approximation result bounding the discretization error in Lemma 3.

Lemma 2. *Let $x \in C_{1-var}([s,t],\mathbb{R}^q)$ and $\mathbf{b} = (b_1,\dots,b_n)$ a window with $b_1 = s$ and $b_n = t$. Let $\delta = t - s$. Then, for any $g \in \mathcal{G}$, there exists some constant $C_{\delta,g}$ depending on $g(x)$ only through $\|g(x)\|_{1-var}$ such that*

$$\|S^I_{\iota,g}(x) - \hat{S}^I_{\iota,g}(x_{\mathbf{b}})\| \le C_{\delta,g}\|g(\exp(x_s))\| \max_{i\in[n-1]} \|g(x)\|_{[b_i,b_{i+1}],1-var}.$$

Proof. Define $z = g(x)$ with the corresponding ι-augmentation denoted by $\tilde{z}$. Then, with $\mathbf{z} = x_{\mathbf{b}}$, by construction, we have $\tilde{\mathbf{z}} = \tilde{z}_{\mathbf{b}}$ (see also Remark 9). We see immediately from our definition of the I-augmented signature that

$$\|S^I_{\iota,g}(x) - \hat{S}^I_{\iota,g}(x_{\mathbf{b}})\| \le \|\exp(g(x_s))\|\|S(\tilde{z}) - S(\tilde{z}_{\mathbf{b}})\|.$$

Appealing to Remark 8 we know that $\|\tilde{z}\|_{1-var} \le C(\|z\|_{1-var} + \|\iota\|_{1-var})$ for some generic constant $C \ge 1$. Since ι is linear between every two points b_i and b_{i+1} we find that $\|\iota\|_{[b_i,b_{i+1}],1-var} = b_{i+1} - b_i$ and $\|\iota\|_{1-var} = \sum_i |b_{i+1} - b_i| = \delta$. But from Corollary 4.3 in Király and Oberhauser [2019], we then find that

$$\begin{aligned}
\|S(\tilde{z}) - S(\tilde{z}_{\mathbf{b}})\| &\le \|\tilde{z}\|_{1-var} e^{\|\tilde{z}\|_{1-var}} \max_{i\in[n-1]} \|\tilde{z}\|_{[b_i,b_{i+1}],1-var} \\
&\le C^2\left(\|z\|_{1-var} + \delta\right) e^{C(\|z\|_{1-var}+\delta)} \max_{i\in[n-1]} \left(\|z\|_{[b_i,b_{i+1}],1-var} + |b_{i+1} - b_i|\right) \\
&\le C_{\delta,g} \max_{i\in[n-1]} \|z\|_{[b_i,b_{i+1}],1-var}
\end{aligned}$$

which, when combined with the previous inequality, yields the statement of the lemma. □

Lemma 3. *Assume that $x, \mathcal{G}$, and $\tau > 0$ satisfy Assumption 2. For any $\epsilon > 0$ and $m \ge 1$, there exists some $\delta > 0$ such that, for any window $\mathbf{b}$ of length τ with $\|\mathbf{b}\| < \delta$, we have*

$$\sup_{g\in\mathcal{G}} \left\| \pi^{(m)} \circ \left(\mathbb{E}S^I_{\iota,g}(x_{[\mathbf{b}]}) - \mathbb{E}\hat{S}^I_{\iota,g}(x_{\mathbf{b}}) \right) \right\| < \epsilon.$$

Proof. Let $\mathbf{b} = (b_1,\dots,b_n)$ be some window such that $b_n = b_1 + \tau$. We define $\varrho(r,s) := \|g(x)\|_{[r,s],1-var}$ so that, by Lemma 2, with $C_t > 0$ a constant depending only on τ and $\|g(x)\|_{[t,t+\tau],1-var}$, it holds that

$$\|\pi^{(m)} \circ \left(S^I_{\iota,g}(x_{[\mathbf{b}]}) - \hat{S}^I_{\iota,g}(x_{\mathbf{b}}) \right)\| \le \sup_{0\le t} C_t \|\pi^{(m)} \circ \exp(g(x_t))\| \max_{i\in[n-1]} \varrho(b_i, b_{i+1}).$$

By way of (3.B.2), for any $\epsilon > 0$, we can find some $\delta > 0$, such that $\max_i \varrho(b_i, b_{i+1}) < \epsilon$ as long as $\|\mathbf{b}\| < \delta$ with the inequality holding uniformly over $\mathcal{G}$ and for almost every path. Additionally, appealing to Hölder's inequality and (3.B.3), we can find some $C' > 0$ such that the expectation of $\sup_{g\in\mathcal{G}} \sup_{0\le t} C_t \|\pi^{(m)} \circ \exp(x_t)\|$ is bounded by C'. Putting all of this together, we find that

$$\mathbb{E}\left(\sup_{g\in\mathcal{G}} \sup_{\mathbf{b}:\|\mathbf{b}\|<\delta} \|\pi^{(m)} \circ \left(S^I_{\iota,g}(x_{[\mathbf{b}]}) - \hat{S}^I_{\iota,g}(x_{\mathbf{b}})\right)\|\right) < C'\epsilon$$

from which the statement of the lemma readily follows. □

3.B.3 Finite sample error

First, a continuity result regarding $\hat{S}^I_{\iota,g}$ and $S^I_{\iota,g}$ when viewed as functions of $g \in \mathcal{G}$.

Lemma 4. *Consider $x \in C_{1-var}([s,t], \mathbb{R}^p)$ and $\mathbf{x} = x_{\mathbf{b}}$ for some window $\mathbf{b} \subset [s,t]$. Let $\mathcal{G} \subset C(\mathcal{O}, \mathbb{R}^p)$ be compact. Then, for all $m \ge 1$,*

(i) if x, $\mathcal{G}$ and $\tau := t - s > 0$ satisfy Eq. (3.B.3), for any $\iota : [s,t] \to [0,1]$ continuous and increasing, the map $\mathcal{G} \ni g \mapsto \pi^{(m)} \circ S^I_{\iota,g}(x)$ is continuous on a set of probability 1,

(ii) for any $\iota \in \Delta_n([0,1])$, on a set of probability 1, the map $\mathcal{G} \ni g \mapsto \pi^{(m)} \circ \hat{S}^I_{\iota,g}(\mathbf{x})$ is continuous.

Proof. Let $(g_n) \subset \mathcal{G}$ converging to $g_\infty \in \mathcal{G}$. Now, note that by Proposition 7.66 in Friz and Victoir [2010], for any $\eta \in [1,2)$,[11] there exists some constant $C_1 > 0$ depending on m, $\sup_{g\in\mathcal{G}} \|g(x)\|_{1-var}$ and $\sup_{g\in\mathcal{G}} \|\pi^{(m)} \circ \exp(g(x_s))\|$ such that

$$\begin{aligned}\left\|\pi^{(m)} \circ \left(S^I_{\iota,g_\infty}(x) - S^I_{\iota,g_n}(x)\right)\right\| \\ \le C_1 \left(\left\|\pi^{(m)} \circ (\exp(g_\infty(x_s)) - \exp(g_n(x_s)))\right\| + \|g_\infty(x) - g_n(x)\|_{\eta-var}\right).\end{aligned}$$

Next, appealing to Proposition 5.5 in Friz and Victoir [2010], we have

$$\begin{aligned}\|g_\infty(x) - g_n(x)\|_{\eta-var} &\le \left(\|g_\infty(x) - g_n(x)\|_{1-var}\right)^{1/\eta} \left(\|g_\infty(x) - g_n(x)\|_\infty\right)^{1-1/\eta} \\ &\le \left(2 \sup_{g\in\mathcal{G}} \|g(x)\|_{1-var}\right)^{1/\eta} \left(\|g_\infty(x) - g_n(x)\|_\infty\right)^{1-1/\eta}.\end{aligned}$$

Now, It follows from (3.B.3) that, on a set of probability 1, both $\sup_{g\in\mathcal{G}} \|\pi^{(m)} \circ \exp(g(x_s))\|$ and $\sup_{g\in\mathcal{G}} \|g(x)\|_{1-var}$ are finite. This immediately implies that, on this set, also C_1 is finite. Thus, part (i) follows if we can show that, for almost every value of the path x, $\|g_\infty(x) - g_n(x)\|_\infty$ goes to 0 for $n \to \infty$. But this follows from the fact that the

[11] For the sake of being precise, we note that the referenced proposition only covers the case where $\eta = 1$. However, the proof works just as well for any $\eta \in (1,2)$.

trace $\mathrm{tr}(x) = \bigcup_{s\leq r\leq t}\{x_r\}$ is almost surely compact (cf. Lemma C.1.(ii) in Schell and Oberhauser [2023]) and g_n converges to g_∞ uniformly on compacts.

For part (ii), by Lemma 2 and Remark 8, and then Lemma 1.(ii), there is some constant $C_2 > 0$ depending only on m, $\sup_{g\in\mathcal{G}} \|g(\mathbf{x})\|_\infty$ and $\sup_{g\in\mathcal{G}} \|\pi^{(m)} \circ \exp(g(x_s))\|$ such that

$$\left\|\pi^{(m)} \circ \left(\hat{S}^I_{\iota,g_\infty}(\mathbf{x}) - \hat{S}^I_{\iota,g_n}(\mathbf{x})\right)\right\| \leq C_2\|g_\infty(\mathbf{x}) - g_n(\mathbf{x})\|_\infty \leq C_2\|g_\infty(x) - g_n(x)\|_\infty.$$

Since $\mathcal{G}$ is compact, we have that C_2 is finite on a set of measure 1 for which, by the same argument as above, we then have the the right-hand side above goes to 0 for $n \to \infty$. This completes the proof. □

Lemma 5. *Let $\mathbf{x}_n = (x_t)_{t\in\mathcal{T}_n}$ be a strongly mixing time-series with x, $\mathcal{G}$ and $\tau > 0$ satisfying Assumption 2 and $\mathcal{G}$ compact. Then, if $\mathfrak{b}_n$ is a batch of size $B_n \to \infty$, for any $\epsilon > 0$, there exists some $n_0 \geq 1$ such that, for all $m \geq 1$ and $n \geq n_0$,*

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}} \|\mathbb{E}\hat{\mathfrak{s}}_m(g(\mathbf{x}_n);\mathfrak{b}_n) - \hat{\mathfrak{s}}_m(g(\mathbf{x}_n);\mathfrak{b}_n)\| > \epsilon\right) < \epsilon.$$

Proof. Let $l : \bigoplus_{k=0}^m (\mathbb{R}^q)^{\otimes k} \to \mathbb{R}$ be linear and define the map $\xi : (D_x)^\infty \times \mathcal{G} \to \mathbb{R}$ given by $\xi : (\mathbf{x}, g) \mapsto l \circ \pi^{(m)} \circ \hat{S}^I_{\iota,g}(\mathbf{x})$.[12] It then suffices to show that

$$\sup_{g\in\mathcal{G}} \|\frac{1}{B_n}\sum_{\mathbf{b}\in\mathfrak{b}_n} \xi(x_\mathbf{b}, g) - \mathbb{E}\xi(x_\mathbf{b}, g)\| \to 0$$

in probability as $n \to \infty$. We shall first prove pointwise convergence (i.e., for a fixed $g \in \mathcal{G}$) and then extend the result to hold uniformly over all of $\mathcal{G}$ using familiar arguments from empirical process theory.

For the first part, fix some $g \in \mathcal{G}$. By a similar argument to the one applied in Lemma F.7 in Schell and Oberhauser [2023], we find that the time series $(\xi(x_\mathbf{b}, g))_{\mathbf{b}\in\mathfrak{b}_n}$ is strongly mixing since $\mathbf{x}_n$ is strongly mixing. Pointwise convergence then follows by the Law of Large Numbers for strongly mixing time series, e.g., Theorem 7.15 in Van der Vaart [2010].

We are then done if we can show that the *bracketing number* (see, for example, Section 6 in Wellner [2005] for a definition) $N_{[]}(\epsilon, \Xi, \|\cdot\|_\infty)$ is finite where $\Xi := \{\xi(\cdot, g) \mid g \in \mathcal{G}\}$. But this follows immediately from Lemma 6.1 in Wellner [2005] noting that, by Lemma 4, the map $g \mapsto \xi(\mathbf{x}, g)$ is continuous for all $\mathbf{x}$, that, by Lemma 3 and (3.B.4), ξ is uniformly bounded, and, finally, that $\mathcal{G}$ is compact by Assumption 2. □

3.B.4 Proof of Lemma 2

For the purpose of readability, we recall the statement of Lemma 2.

[12] Here $(D_x)^\infty$ denotes the space of all finite sequences with elements in D_x. We note that ξ is well defined by first identifying an element $\mathbf{x} \in (D_x)^\infty$ with the corresponding vector in $(D_x)^l$ (where l is the smallest integer such that x_j is 0 for all $j > l$) for which $\hat{S}^I_{\iota,g}(\mathbf{x})$ is defined as per usual.

Lemma 6 (Lemma 2)**.** *Let $\mathcal{G} \subset C(\mathbb{R}^p)$ compact and $\mathbf{x}_n = (x_t)_{t\in\mathcal{T}_n}$ with the time grids $\mathcal{T}_n$ as described in the beginning of Section 3.4.1. Assume that x, $\mathcal{G}$ and $\tau > 0$ satisfy the technical Assumption 2. Assume that $\mathbf{x}_n$ is strongly mixing. Let $\mathfrak{b}_n$ be a sequence batches of size B_n with $B_n \to \infty$ as n increases and such that each window is of length $\tau > 0$. Then, for any $\epsilon > 0$, there exists $m_0 \geq 1$ such that, for all $m \geq m_0$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{g\in\mathcal{G}} \|\hat{\mathfrak{s}}_m(g(\mathbf{x}_n);\mathfrak{b}_n) - \mathfrak{s}(g(x);\mathfrak{b}_n)\| > \epsilon\right) \to 0.$$

Proof. Fix some $\epsilon > 0$. To prove the first part we start with (3.B.1) and bound each term seperately. Let us call the respective terms (a), (b), and (c), i.e.,

$$\begin{aligned}
(a) &:= \sup_{g\in\mathcal{G}} \left\|\mathfrak{s}(g(x);\mathfrak{b}_n) - \pi^{(m)} \circ \mathfrak{s}_m(g(x);\mathfrak{b}_n)\right\|, \\
(b) &:= \sup_{g\in\mathcal{G}} \left\|\pi^{(m)} \circ \mathfrak{s}(g(x);\mathfrak{b}_n) - \mathbb{E}\hat{\mathfrak{s}}_m(g(\mathbf{x}_n);\mathfrak{b}_n)\right\| \\
(c) &:= \sup_{g\in\mathcal{G}} \|\mathbb{E}\hat{\mathfrak{s}}_m(g(\mathbf{x}_n);\mathfrak{b}_n) - \hat{\mathfrak{s}}_m(g(\mathbf{x}_n);\mathfrak{b}_n)\|.
\end{aligned}$$

(a): By Lemma 1, we can find some $m_0 \geq 1$ such that, for all $m \geq m_0$ and windows $\mathbf{b}$ of length $\tau > 0$, it holds that $\sup_{g\in\mathcal{G}} \left\|\mathbb{E}S^I_{\iota,g}(x_{[\mathbf{b}]}) - \mathbb{E}\left(\pi^{(m)} \circ S^I_{\iota,g}(x_{[\mathbf{b}]})\right)\right\| < \epsilon$ which implies that

$$(a) \leq \frac{1}{B_n} \sum_{\mathbf{b}\in\mathfrak{b}_n} \sup_{g\in\mathcal{G}} \left\|\mathbb{E}S^I_{\iota,g}(x_{[\mathbf{b}]}) - \mathbb{E}\left(\pi^{(m)} \circ S^I_{\iota,g}(x_{[\mathbf{b}]})\right)\right\| < \frac{\epsilon}{3}.$$

(b): Now, let $m \geq m_0$ with m_0 as above and take $\delta > 0$ such that

$$\sup_{g\in\mathcal{G}} \left\|\pi^{(m)} \circ \left(\mathbb{E}S^I_{\iota,g}(x_{[\mathbf{b}]}) - \mathbb{E}\hat{S}^I_{\iota,g}(x_{\mathbf{b}})\right)\right\| < \frac{\epsilon}{3}$$

for all windows $\mathbf{b}$ with $\|\mathbf{b}\| < \delta$. We note that such a choice is possible due to Lemma 3. Since we have assumed that the mesh of the time grid $\mathcal{T}_n$ goes to zero and each window in $\mathfrak{b}_n$ is of fixed length τ, we must have $\|\mathbf{b}\| \to 0$ for all $\mathbf{b} \in \mathfrak{b}_n$. Then, pick $n_0 \geq 1$ large enough so that $\max_{\mathbf{b}\in\mathfrak{b}_n} \|\mathbf{b}\| < \delta$ for all $n \geq n_0$ and, hence,

$$(b) \leq \frac{1}{B_n} \sum_{\in\mathfrak{b}_n} \sup_{g\in\mathcal{G}} \left\|\pi^{(m)} \circ \left(\mathbb{E}S^I_{\iota,g}(x_{[\mathbf{b}]}) - \mathbb{E}\hat{S}^I_{\iota,g}(x_{\mathbf{b}})\right)\right\| < \frac{\epsilon}{3}.$$

(c): It follows directly from Lemma 5 that, for any $m \geq m_0$, we can pick $n_1 \geq n_0$ large enough so that $\mathbb{P}((c) > \epsilon/3) < \epsilon$. Combining all three parts, we then obtain (for our specific choice of $m \geq m_0$) that, for all $n \geq n_1$,

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}} \|\hat{\mathfrak{s}}_m(g(\mathbf{x}_n);\mathfrak{b}_n) - \mathfrak{s}(g(x);\mathfrak{b}_n)\| > \epsilon\right) \leq \mathbb{P}\left((a) + (b) + (c) > \epsilon\right) \leq \mathbb{P}\left((c) > \frac{\epsilon}{3}\right) < \epsilon$$

which completes the proof. □

3.B.5 Proof of Theorem 3

For convenience, we define the intermediate version of the oracle, $\tilde{\varphi}_{sig}$, given by

$$\tilde{\varphi}_{sig}(g(x); \mathfrak{B}_n) := B_n^{-2} \sum_{\mathfrak{b},\mathfrak{b}' \in \mathfrak{B}_n} \|\mathfrak{s}(g(x); \mathfrak{b}) - \mathfrak{s}(g(x); \mathfrak{b}')\|^2.$$

In essence, the aim of the above was to prove that $\hat{\varphi}_{sig}$ approximates $\tilde{\varphi}_{sig}$ uniformly over $\mathcal{G}$. Together with Assumption 1, the proof of Theorem 3 then follows easily. Recall that, for $e = (e_1, e_2)$ the ground truth stationary embedding and $[e]_{\mathcal{F}}$ the identifiable equivalence class of stationary embeddings, we define $[e_1]_{\mathcal{G}}$ as the set of all maps $g_1 \in \pi_{1:k}([e]_{\mathcal{F}})$ such that $(g_1, g_2) := g \in \mathcal{G}$ for some suitable g_2.

Theorem 4 (Theorem 3)**.** *Let x be $\mathcal{CI}_k$ with $x = d(y, z)$ and $\mathbf{y}_n = (y_t)_{t \in \mathcal{T}_n}$ strongly mixing with the time grids $\mathcal{T}_n$ as described in the beginning of Section 3.4.1. For $\mathcal{F}$ a class of admissible mixings with $d \in \mathcal{F}$ and $\mathfrak{B}_n$ a batching protocol, assume that $(\mathcal{F}, z)$ is identifiable with z satisfying Assumption 1 for $\tau > 0$ and every $\mathfrak{b}, \mathfrak{b}'_n \in \mathfrak{B}_n$. Finally, for $\mathcal{G} \subset C(\mathcal{O}, \mathbb{R}^p)$, assume that $x, [e_1]_{\mathcal{G}}$, and τ satisfy the technical Assumption 2. Then, for any $\epsilon > 0$, there is a sufficiently large truncation level $m \geq 1$ such that, for $\hat{e}_1$ as in (3.4.6), we have*

$$\lim_{n \to \infty} \mathbb{P}\left(\mathrm{d}_\infty\left(\hat{e}, [e]_{\mathcal{F}}\right) > \epsilon\right) = 0.$$

Proof. Throughout, we let $\mathcal{G}_1 := \{g_{1:k} \mid g \in \mathcal{G}\}$. For some $\epsilon > 0$ small enough we let $[e_1]^{\epsilon}_{\mathcal{G}} \subset \mathcal{G}_1$ be given by

$$[e_1]^{\epsilon}_{\mathcal{G}} := \bigcup_{g \in [e_1]_{\mathcal{G}}} B_\epsilon(g).$$

By Assumption 1, there exists some $c_\epsilon > 0$ and $n_0 \geq 1$ such that $\mathbb{P}(\hat{\varphi}_{sig}(g(\mathbf{x}_n); \mathfrak{B}_n) \leq c_\epsilon) < \epsilon$ for all $g \in \mathcal{G}_1 \setminus [e_1]^{\epsilon}_{\mathcal{G}}$ and $n \geq n_0$. Now take $m_0 \geq 1$ and $n_1 \geq n_0$ (the latter depending on m_0) large enough so that, for all $n \geq n_1$,

$$\mathbb{P}\left(\min_{g \in \mathcal{G}_1} \hat{\varphi}_{sig}(g(\mathbf{x}_n); \mathfrak{B}_n) \geq c_\epsilon\right) < \epsilon.$$

Such a choice is possible due to Lemma 2 since $\tilde{\varphi}_{sig}(g(x)) = 0$ and $g(x)$ is strongly mixing for any $g \in [e_1]_{\mathcal{G}}$. Let B_n denote complement of the event inside the probability. Combining these two inequalities, we find that

$$\begin{aligned}
\mathbb{P}\left(\pi_{1:k} \circ \hat{e} \notin [e_1]_{\mathcal{G}}\right) &\leq \mathbb{P}\left(\hat{\varphi}_{sig}(\pi_{1:k}(\mathbf{x}_n); \mathfrak{B}_n) \leq c_\epsilon \mid \pi_{1:k} \circ \hat{e} \notin [e_1]_{\mathcal{G}}\right) \\
&\quad + \mathbb{P}\left(\hat{\varphi}_{sig}(\pi_{1:k}(\mathbf{x}_n); \mathfrak{B}_n) \geq c_\epsilon\right) \\
&\leq 2\epsilon.
\end{aligned}$$

Thus, for all $n \geq n_1$, the probability that $\mathrm{d}_\infty(\hat{e}, [e]_{\mathcal{F}}) < \epsilon$ is bounded from below by $1 - 2\epsilon$. This completes the proof. □

3.B.6 Remarks on Assumption 1

As mentioned in the main body of text, Assumption 1 is a high level assumption and not easily interpretable. The main difficulty will be to find classes of processes for which this assumption holds. This is the purpose of the present section. Throughout, we shall consider cases where also the non-stationary component is strongly mixing and satisfies the technical Assumption 2. In particular, we consider the following setup: Let x_t be $\mathcal{CI}_k$ with $x_t = d(y_t, z_t)$ and $\mathbf{x}_n = (x_t)_{t\in\mathcal{T}_n}$ strongly mixing. For $\mathcal{F}$ a class of admissible mixings with $d \in \mathcal{F}$ and $\mathfrak{B}_n$ a batching protocol, assume that $(\mathcal{F}, z)$ is identifiable with z_t satisfying Assumption 1 for $\tau > 0$ and every $\mathfrak{b}, \mathfrak{b}'_n \in \mathfrak{B}_n$. Finally, for $\mathcal{G} \subset C(\mathcal{O}, \mathbb{R}^p)$, assume that $x, \mathcal{G}$, and τ satisfy the technical Assumption 2. Lemma 2 then ensures that the difference between $\hat{\varphi}_{sig}$ and $\tilde{\varphi}_{sig}$ converges to 0 in probability as $n \to \infty$ uniformly over $\mathcal{G}_1$.

Example 3.6.1 (Piece-wise stationary)**.** Consider two sequences of open sets $\mathcal{I}_n, \mathcal{J}_n \subset [0,\infty)$ such that $\mathcal{I}_n \cap \mathcal{J}_n = \emptyset$ for all n. For any open $U \subset [0,\infty)$, let $\mathcal{N}_\tau(U)$ denote the largest number $N \geq 0$ such we can fit N disjoint intervals of length τ into U. Now, assume that $\mathcal{N}(\mathcal{I}_n), \mathcal{N}(\mathcal{J}_n) \to \infty$ for $n \to \infty$ and that for any two pairs of disjoint $(s_1, s_1+\tau) \in \mathcal{I}_{n_1}, (t_1, t_1+\tau) \in \mathcal{I}_{n_2}$ and $(s_2, s_2+\tau) \in \mathcal{J}_{n_3}, (t_2, t_2+\tau) \in \mathcal{J}_{n_4}$ the distributions of $z_{[s_i, s_i+\tau]}$ and $z_{[t_i, t_i+\tau]}$ agree for $i = 1, 2$ but the distributions of $z_{[s_1, s_1+\tau]}$ and $z_{[s_2, s_2+\tau]}$ differ. In other words, there are two growing regions on which the z_t is stationary, however, the distribution of z_t changes between these regions. An example of such a process is given in Example 3.5.1. Then, Assumption 1 is satisfied. To see this, first note that, since z_t is strictly non-stationary, for any $g \notin [e_1]_\mathcal{G}$, we will have that $g(x_t)$ is piecewise stationary and non-stationary with the regions $\mathcal{I}_n$ and $\mathcal{J}_n$ as above. In particular, for any two windows $\mathbf{b}_i \in \mathcal{I}_n$ and $\mathbf{b}_j \in \mathcal{J}_n$, we have

$$\left\| \mathbb{E} S^I_{\iota,g}(x_{[\mathbf{b}_i]}) - \mathbb{E} S^I_{\iota,g}(x_{[\mathbf{b}_j]}) \right\|^2 := c_g > 0.$$

Now define two batches $\mathfrak{b}^\mathcal{I}_n$ and $\mathfrak{b}^\mathcal{J}_n$ such that $\mathfrak{b}^\mathcal{I}_n$ (resp. $\mathfrak{b}^\mathcal{I}_n$) consists of $\mathcal{N}(\mathcal{I}_n)$ (resp. $\mathcal{N}(\mathcal{J}_n)$) disjoint windows of length $\tau > 0$ in $\mathcal{I}_n$ (resp. $\mathcal{J}_n$). With $\mathfrak{B}_n = (\mathfrak{b}^\mathcal{I}_n, \mathfrak{b}^\mathcal{J}_n)$, it is then not hard to see that $\tilde{\varphi}_{sig}(g(x); \mathfrak{B}_n) = c_g > 0$. With $\tilde{\varphi}^m_{sig}$ defined analogously to $\tilde{\varphi}_{sig}$, but with the full signature replaced by the m-th order truncation, Lemma 3.B.1 then yields, for all $g \notin [e_1]_\mathcal{G}$,

$$\lim_{m\to\infty} \tilde{\varphi}^m_{sig}(g(x); \mathfrak{B}_n) = \tilde{\varphi}_{sig}(g(x); \mathfrak{B}_n).$$

Now, by Lemma 4, $\tilde{\varphi}^m_{sig}$ is continuous as a map of g for every $m \geq 1$ so that $\lim_{n\to\infty} \tilde{\varphi}^m_{sig}$ is lower semi-continuous on $\mathcal{G}_1$. But then, by compactness, for any $\epsilon > 0$ small enough, we have

$$\inf_{g \in \mathcal{G}_1 \setminus [e_1]^\epsilon_\mathcal{G}} \tilde{\varphi}_{sig}(g(x); \mathfrak{B}_n) := c_\epsilon > 0$$

where the constant c_ϵ depends only on ϵ and not on n. The result then follows upon realizing that, by Lemma 2, $\tilde{\varphi}_{sig}$ uniformly approximates $\hat{\varphi}_{sig}$ in probability over $\mathcal{G}_1$. ♠

3.C. Signatures

In this part of the appendix we introduce signatures and the corresponding kernel. Similar to Király and Oberhauser [2019], we define a separate object for sequences which we call the discrete signature. Much of the following is standard in the literature with a few new ideas added along the way. We refer the interested reader to Lee and Oberhauser [2023], Cass and Salvi [2024] for more information on signatures and their applications in machine learning.

Throughout, we let H be some generic real Hilbert space (e.g., $H = \mathbb{R}^p$). We define the *extended tensor algebra* over H, also denoted $T((H))$, to be the space of all formal sequences $\mathbf{a} = (\mathbf{a}_0, \mathbf{a}_1, \dots)$ where $\mathbf{a}_k \in H^{\otimes k}$ for all $k \geq 0$ and we use the convention $H^{\otimes 0} = \mathbb{R}$. We define addition and multiplication on $T((H))$ as follows: For any two elements $\mathbf{a}, \mathbf{b} \in T((H))$ the sum $\mathbf{c} = \mathbf{a} + \mathbf{b}$ is just the formal series obtained by pointwise addition and the product $\mathbf{c} = \mathbf{ab}$ is the new formal series where the k'th element is given by $\sum_{0 \leq j \leq k} \mathbf{a}_j \otimes \mathbf{b}_{k-j}$. The extended tensor algebra is then a real non-commutative unital algebra. We observe that each $H^{\otimes k}$ can naturally be viewed as a Hilbert space by completion under the natural inner product

$$\langle f_1 \otimes \cdots \otimes f_k, g_1 \otimes \cdots \otimes g_k \rangle_{H^{\otimes k}} = \prod_{1 \leq j \leq k} \langle f_j, g_j \rangle_H.$$

Similarly, we can then view $T((H))$ as a Hilbert space equipped with the inner product $\langle \mathbf{a}, \mathbf{b} \rangle_{T((H))} := \sum_{k \geq 0} \langle \mathbf{a}_k, \mathbf{b} \rangle_{H^{\otimes k}}$. The *truncated tensor algebra of order* $m \geq 1$, also denoted $T^{(m)}(H)$, is the algebra given by $\bigoplus_{k=0}^{m} H^{\otimes k}$ with addition and multiplaction defined the same as above. For each $m \geq 1$, we denote the canonical projection of $T((H))$ onto $T^{(m)}(H)$ by $\pi^{(m)}$.

Recall that $C_{1-var}([s,t], H)$ denotes the space of continuous paths form $[s,t]$ into H with finite 1-variation (see also Defintion 12). The (continuous) *signature* is a map $S : C_{1-var}([s,t], H) \to T((H))$ given by $S(x) = (1, S_1(x), S_2(x), \dots)$ with $S_k(x) \in H^{\otimes k}$ satisfying

$$S_k(x) = \int \cdots \int_{s < u_1 < \cdots < u_k < t} dx_{u_1} \otimes \cdots \otimes dx_{u_k}. \tag{3.C.1}$$

The *truncated signature of order* $m \geq 1$ is simply the projection of the signature onto the truncated tensor algebra of order m, i.e., $\pi^{(m)} \circ S$. Finally, the *signature kernel* is the kernel on $C_{1-var}([s,t], H)$ given by $k(x,y) = \langle S(x), S(y) \rangle_{T((H))}$. The truncated signature kernel is then defined analogously for any $m \geq 1$.

3.C.1 Variation norms

The following definition of the η-variation corresponds to the usual definition of what is most commonly referred to as p-variation in the literature. We use a slightly different terminology to avoid overloading the notation.[13] We also define the natural adaptation of η-variation for discrete sequences.

[13] Note that throughout we are using p to denote the dimension of the observable sample space.

Definition 12 (η-variation)**.** Let $\eta \geq 1$. For a sequence $\mathbf{x} = (x_1, \dots, x_n) \in H^n$, we define the *$\eta$-variation* of x as

$$\|\mathbf{x}\|_{\eta-var} := \max_{k\in[n]} \max_{\mathbf{i}\in\Delta_k([n])} \left(\sum_{j\in[k-1]} \|\Delta x_{\mathbf{i}_j}\|^\eta \right)^{1/\eta}.$$

For a continuous path $x \in C([s,t], H)$, we define the *η-variation* of x as

$$\|x\|_{[s,t],\eta-var} := \sup_{\mathcal{T}\subset[s,t]} \|x_{\mathcal{T}}\|_{\eta-var}$$

where the supremum is taken over all *partitions* of $[s,t]$.[14] We shall sometimes omit the subscript signifying the dependence on the interval $[s,t]$ when this does not otherwise cause confusion.

Lemma 1. *Let $x \in C([s,t], H)$ and $\mathbf{x} \in H^n$. The following holds:*

(i) Discretisation decreases the η-variation, i.e., if $\mathbf{x} = (x_r)_{r\in\mathcal{T}}$ for some time grid $\mathcal{T} = (s \leq r_1 < \cdots < r_k \leq t)$, then

$$\|\mathbf{x}\|_{\eta-var} \leq \|x\|_{\eta-var}.$$

(ii) There exists some $C_n > 0$ only depending on n, such that

$$\|\mathbf{x}\|_{1-var} \leq C_n \|\mathbf{x}\|_\infty.$$

(iii) For continuous paths, the definition agrees with the usual definition, i.e.,

$$\|x\|_{\eta-var} = \sup_{\mathcal{T}\subset[s,t]} \left(\sum_{(u,v)\in\mathcal{T}} \|x_v - x_u\|^\eta \right)^{1/\eta},$$

where the supremum is taken over all partitions of $[s,t]$.

Proof. The first part follows directly from the definition of η-variation noting simply that $\mathbf{x} = x_{\mathcal{T}}$.

The second part follows immediately from the observation that, for any $k \in [n]$ and $\mathbf{i} \in \Delta_k([n])$,

$$\sum_{j\in[k-1]} \|\Delta x_{\mathbf{i}_j}\| \leq 2n\|\mathbf{x}\|_\infty.$$

[14] A partition of $[s,t]$ is a finite set of numbers $P = (t_0 < \cdots < t_n\}$ such that $t_0 = s$ and $t_n = t$. This is slightly different from what we have called a time grid where the starting and end point need not align with s and t, i.e., every partition is a time grid, but not vice-versa.

For the third part, we observe that, since $\Delta_n([n]) = [n]$, the "$\geq$"-direction is trivial. To show the other direction, let $\mathcal{T}_n \subset [s,t]$ be some partition with n points. We write $\mathcal{T}_n = (s = r_1 < \cdots < r_n = t)$. For some $k \in n$ and $\mathbf{i} \in \Delta_k([n])$, let $\mathcal{T}_k \subset [s,t]$ be the partition of at most $k+2$ points that agrees with $(r_{\mathbf{i}_1}, \dots, r_{\mathbf{i}_k})$ up to adding s or t at either end if $\mathbf{i}_1 > 1$ or $\mathbf{i}_k < n$. Then,

$$\sum_{j \in [k-1]} \|\Delta x_{r_{\mathbf{i}_j}}\|^\eta \leq \sum_{(u,v) \in \mathcal{T}_k} \|x_u - x_v\|^\eta \leq \sup_{\mathcal{T} \subset [s,t]} \sum_{(u,v) \in \mathcal{T}} \|x_u - x_v\|^\eta.$$

The right hand side does not depend on $\mathcal{T}_n$, k, or $\mathbf{i}$. Therefore, taking the supremum over all such choices we obtain the other inequality and the proof is done. □

3.C.2 Discrete Signature

There are multiple ways to extend the continuous signature to cover sequences. Perhaps the most natural way is to first embed the sequence in the space of paths of bounded 1-variation, e.g., by taking the linear interpolation, after which one can then use the standard definition of the signature. Here we shall adapt the approach of Király and Oberhauser [2019] and instead approximate the iterated integrals in (3.C.1) directly with iterated sums. Our main reasons for doing so are exclusively computational. See also Appendix 3.D.

Definition 13 (Discrete signature)**.** Fix some $n \geq 0$. The *discrete signature* is defined as the map $\hat{S} : H^n \to T((H))$ given by $\hat{S}(\mathbf{x}) = (1, \hat{S}_1(\mathbf{x}), \dots)$ where, for $k \leq n-1$,

$$\hat{S}_k(\mathbf{x}) := \sum_{\mathbf{i} \in \Delta_k([n-1])} \Delta\mathbf{x}_{\mathbf{i}_1} \otimes \cdots \otimes \Delta\mathbf{x}_{\mathbf{i}_k}$$

and $\hat{S}_k(\mathbf{x}) = 0 \in H^{\otimes k}$ for all $k \geq n$. Here $\Delta\mathbf{x}_i := \mathbf{x}_{i+1} - \mathbf{x}_i$.

If we let H^∞ denote the space of all finite sequences in H, we can naturally extend the discrete signature to have all of H^∞ as its domain. The *discrete signature kernel*, call it $\hat{k}$, is then the map $\hat{k} : H^\infty \times H^\infty \to \mathbb{R}$ given by $\hat{k}(\mathbf{x}, \mathbf{y}) = \langle \hat{S}(\mathbf{x}), \hat{S}(\mathbf{y}) \rangle_{T((H))}$.

The following bound is quite crude, but it is easy to show and suffices for the present purposes. In particular, it shows that the discrete signature is Lipschitz continuous on sets of bounded variation (provided the input sequences are of equal length).

Lemma 2. *Let $\mathbf{x}, \mathbf{y} \in H^n$ for $n \geq m \geq 1$. Then, there exists a constant $C_{n,m}$ depending on x and y only through $\|\mathbf{x}\|_{1-var}$ and $\|\mathbf{y}\|_{1-var}$, such that*

$$\left\| \pi^{(m)} \circ \left(\hat{S}(\mathbf{x}) - \hat{S}(\mathbf{y}) \right) \right\| \leq C_{n,m} \|\mathbf{x} - \mathbf{y}\|_{1-var}. \tag{3.C.2}$$

Proof. By definition of the discrete signature, we observe that

$$\begin{aligned}
\left(\hat{S}(\mathbf{x}) - \hat{S}(\mathbf{y}) \right) &= \left(\sum_{\mathbf{i} \in \Delta_k([n])} \Delta x_{\mathbf{i}_1} \otimes \cdots \otimes \Delta x_{\mathbf{i}_k} - \Delta y_{\mathbf{i}_1} \otimes \cdots \otimes \Delta y_{\mathbf{i}_k} \right)_{k \geq 0} \\
&= \left(\sum_{\mathbf{i} \in \Delta_k([n])} \sum_{j \in [k]} \Delta x_{\mathbf{i}_1} \otimes \cdots \otimes \Delta x_{\mathbf{i}_{j-1}} \otimes \Delta(x-y)_{\mathbf{i}_j} \otimes \Delta y_{\mathbf{i}_{j+1}} \otimes \cdots \otimes \Delta y_{\mathbf{i}_k} \right)_{k \geq 0}
\end{aligned}$$

from which it follows that

$$\left\|\pi^{(m)} \circ \left(\hat{S}(\mathbf{x}) - \hat{S}(\mathbf{y})\right)\right\| \leq \sum_{k\in[m]} \sum_{\mathbf{i}\in\Delta_k([n])} \sum_{j\in[k]} \|\Delta x_{\mathbf{i}_1}\| \cdots \|\Delta x_{\mathbf{i}_{j-1}}\| \|\Delta(x-y)_{\mathbf{i}_j}\| \|\Delta y_{\mathbf{i}_{j+1}}\| \cdots \|\Delta y_{\mathbf{i}_k}\|$$
$$\leq \|\mathbf{x}-\mathbf{y}\|_{1-var} \sum_{k\in[m]} \sum_{\mathbf{i}\in\Delta_k([n])} \sum_{j\in[k-1]} \|\mathbf{x}\|_{1-var}^{j-1} \|\mathbf{y}\|_{1-var}^{k-j-1}.$$

□

3.C.3 Signature Augmentations

As remarked in the main body of the text, the signature characterizes paths only up to translations by tree-like or constant paths. We shall now define two ways to augment the signature so that it becomes sensitive to either of these two cases. In particular, with Theorem 5 we then obtain a metric for compactly supported probability measures on the space of paths of bounded variation.

Definition 14 (ι-augmentation)**.** Let $x \in C_{1-var}([s,t],H)$ and $\iota : [s,t] \to [0,1]$ continuous and strictly increasing. The *ι-augmentation* of x is the path in $H \times \mathbb{R}$ given by $\tilde{x} = (x, \iota)$. Similarly, for a sequence $\mathbf{x} \in H^n$, we define the *ι-augmentation* for some $\iota \in \Delta_n([0,1])$ as the sequence in $H \times \mathbb{R}$ given by

$$\tilde{\mathbf{x}} = ((x_1,\iota_1),\ldots,(x_n,\iota_n)).$$

Remark 8. It follows immediately from Definition 14 that, for any two $x, y \in C_{1-var}([s,t],H)$, we have $\|\tilde{x}-\tilde{y}\|_{1-var} = \|x-y\|_{1-var}$ where $\tilde{x}$ and $\tilde{y}$ are the respective ι-augmentation for a specific given ι. Furthermore, it is not hard to prove that $\|\tilde{x}\|_{1-var} \leq C(\|x\|_{1-var} + \|\iota\|_{1-var})$ for some constant $C > 0$ not depending on x or ι. Finally, if $\tilde{x}'$ is the augmentation obtained from some alternative ι', then $\|\tilde{x}-\tilde{x}'\|_{1-var} = \|\iota-\iota'\|_{1-var}$. Similar results holds for the ι-augmentation of sequences with the continuous 1-variation replaced by the discrete 1-variation.

Remark 9. Consider $x \in C([s,t],H)$ and $\mathbf{x} := x_{\mathbf{b}} := (x_t)_{t\in\mathbf{b}}$ for some window $\mathbf{b} \in \Delta_n([s,t])$. Then, for $\iota : [s,t] \to [0,1]$ continuous and increasing, we have $(\iota(b))_{b\in\mathbf{b}} \in \Delta_n([0,1])$ and, abusing notation slightly, we shall write $\tilde{x}$ and $\tilde{\mathbf{x}}$ for the respective ι-augmentations. Note that this is perfectly consistent. Indeed, we have $\tilde{x}_{\mathbf{b}} = \tilde{\mathbf{x}}$.

Adding a strictly increasing coordinate to the path x ensures that the signature distinguishes paths up to translation by a constant. Now, there are multiple approaches to make the signature sensitive to translation by constants. The simple idea that we shall follow is to add some information capturing the initial value of the path. Crucially, to show universality of the signature as a feature map, we need to ensure that the image of this augmented signature is still *group-like* (see, e.g., Definition 2.18 in Lyons et al. [2007]). A simple way to achieve this is then simply to pre-multiply the signature by the tensor exponential of the initial value x_s. Recall that the tensor exponential is the map $\exp : H \to T((H))$ such that the k'th element of $\exp(f)$ is given by $f^{\otimes k}/k!$.

Definition 15 (I-augmentaition)**.** The *I-augmentation* of the signature is the map $S^I : C_{1-var}([s,t],H) \to T((H))$ given by $S^I(x) = \exp(x_s)S(x)$. Similarly, the *I-augmentation* of the discrete signature is the map $\hat{S}^I : H^\infty \to T((H))$ given by $\hat{S}^I(\mathbf{x}) = \exp(\mathbf{x}_0)\hat{S}(\mathbf{x})$.

Remark 10. In the continuous case, there is a clear interpretation of the I-augmentation on the level of paths. Indeed, $S^I(x)$ is simply the signature of the path obtained by concatenating the straight line joining the origin to x_s to the original path x. Unfortunately there is no such interpretation in the discrete case. The natural counterpart would by appending a 0 to the beginning of the sequence and then taking the discrete signature. As can be readily checked, however, this would instead result in $\mathbf{x}_0\hat{S}(\mathbf{x})$.

For $\iota : [s,t] \to [0,1]$ continuous and strictly increasing, we shall write S^I_ι for the composition $x \mapsto \tilde{x} \mapsto S^I(\tilde{x})$ where $\tilde{x}$ is the ι-augmentation of x. In the following we equip $C_{1-var}([s,t],H)$ with topology induced by the 1-*variation norm* defined by $\|x\|_1 := \|x_s\| + \|x\|_{1-var}$ under which it is a Banach space (see, e.g., Theorem 1.28 in Friz and Victoir [2010], the proof of which works for general Banach-valued paths).

Lemma 3. *Let $\iota : [s,t] \to [0,1]$ be continuous and strictly increasing. Then, the map $S^I_\iota : C_{1-var}([s,t],H) \to T((H))$ is injective and continuous.*

Proof. Let $x, y \in C_{1-var}([s,t],H)$. We note that

$$\|S^I_\iota(x) - S^I_\iota(y)\| \leq \|\exp(x_s) - \exp(y_s)\|\|S_\iota(x)\| + \|\exp(y_s)\|\|S_\iota(x) + S_\iota(y)\|.$$

Now, the first term is 0 if, and only if, $x_s = y_s$ and the second term is 0 if, and only if, $x = y + c$ for some constant $c \in H$ (see, for example, Proposition 1.2.4 in Cass and Salvi [2024]). This proves injectivity.

For continuity, note that we have that, e.g., by Corollary 5.5 in Chevyrev and Lyons [2016], the signature is continuous in 1-variation. It therefore suffices to show that the tensor exponential is continuous. But this is easy to see. Indeed, since

$$\|x_s^{\otimes k} - y_s^{\otimes k}\| \leq \|x_s\|^k + \|y_s\|^k,$$

the higher order terms of $\|\exp(x_s) - \exp(y_s)\|$ become negligible for $k \geq 1$ large enough. Continuity then follows directly from continuity of the maps $a \mapsto a^{\otimes k}$. □

Theorem 5 (Signature MMD)**.** *Let $K \subset C_{1-var}([s,t],H)$ be compact and $\iota : [s,t] \to [0,1]$ continuous and strictly increasing. Then, for any two $\mu, \nu \in \mathcal{M}_1(K)$, we have $\mu = \nu$ if, and only if,*

$$\mathbb{E}_{x\sim\mu} S^I_\iota(x) = \mathbb{E}_{x\sim\nu} S^I_\iota(x).$$

Proof. We need to show that the map $\mathcal{M}_1(K) \ni \mu \mapsto \mathbb{E}_{x\sim\mu} S^I_\iota(x)$ is injective. This is equivalent to showing (by, e.g., Theorem 7 in Chevyrev and Oberhauser [2022]) that the augmented signature S^I_ι is universal, i.e., that linear functions of the signature are dense in $C(K,\mathbb{R})$. We shall take the usual approach and prove this by an application of the Stone-Weierstrass Theorem. To this end, we define $\Omega := \{l \circ S^I_\iota \mid l \in T((H))^*\}$ where $T((H))^*$ denotes the topological dual of $T((H))$. It follows from Lemma 3 that

$\Omega \subset C(K, \mathbb{R})$. Furthermore, we note that Ω is a subalgebra (see, e.g., section 2.2.3 of Lyons et al. [2007])). Since S_ι^I is also injective (referring again to Lemma 3), we find that Ω separates points. Thus, by the Stone-Weierestrass Theorem, Ω is dense in $C(K, \mathbb{R})$ as desired. □

3.D. Numerics

Before discussing the main two algorithms 3 and 4, we first present a way to approximate the discrete signature kernels of sequences $\kappa_\gamma(\mathbf{x}, \cdot)$ where $\kappa_\gamma(x, y) = \exp(-\|x-y\|^2/2\gamma^2)$ is the Gaussian kernel. The method was recently introduced in Toth et al. [2023] and relies on the fact that any translation invariant kernel can be represented as the Fourier transform of its *spectral measure* (see also Rahimi and Recht [2007]). As such, the same ideas could, in principle, also be applied to other translation invariant kernels such as, for example, the Matérn kernels. We will generally refer to the features obtained via this approach as *random Fourier signature features*. To be precise, we shall approximate the signature kernel using the *diagonally projected random Fourier Signature features* (or RFSF-DP) truncated at some depth $m \geq 1$ as given in Definition 3.4 of Toth et al. [2023]. See also their Theorem 3.5 for concentration guarantees. Crucially, this will then enable us to develop algorithms with run-times complexity scaling linearly in sequence length n and dimension p. For readability, we restate the definition of RFSF-DP here.

Definition 16 (RFSF-DP)**.** Let $\mathbf{x} = (x_1, \dots, x_{n_x})$ and $\mathbf{y} = (y_1, \dots, y_{n_y})$ be sequences in $\mathbb{R}^p$, $\gamma > 0$ the variance of the Gaussian kernel, $m \geq 1$ the depth, and $N_f \geq 1$ and the number of samples to use in the approximation. For $l \in [m]$ and $q \in [N_f]$, draw the i.i.d. p-dimensional weights $w_q^l \sim \mathcal{N}(0, \gamma I_p)$. The *RFSF-DP* map truncated at depth m is then given by

$$\tilde{S}^{(m)}_{N_f,\gamma}(\mathbf{x}) = \frac{1}{\sqrt{N_f}} \left(\left(\sum_{\mathbf{i} \in \Delta_l(n_x-1)} \Delta\tilde{x}^{1,q}_{\mathbf{i}_1} \otimes \cdots \otimes \Delta\tilde{x}^{l,q}_{\mathbf{i}_l} \right)_{q \in [N_f]} \right)_{l \in [m]} \in \bigoplus_{l=0}^{m} \left(\left(\mathbb{R}^2\right)^{\otimes l} \right)^{N_f} := \tilde{\mathcal{H}}^{(m)}_{N_f}$$

where $\tilde{x}_i^{l,q} = (\cos(x_i^\intercal w_q^l), \sin(x_i^\intercal w_q^l)) \in \mathbb{R}^2$ and similarly defined for $\mathbf{y}$. The *RFSF-DP* kernel, call it $\tilde{k}^{(m)}_{N_f,\gamma}$, is then simply computed as the inner product in $\tilde{\mathcal{H}}^{(m)}_{N_f}$, that is,

$$\begin{aligned} \tilde{k}^{(m)}_{N_f,\gamma}(\mathbf{x}, \mathbf{y}) &= \langle \tilde{S}^{(m)}_{N_f,\gamma}(\mathbf{x}), \tilde{S}^{(m)}_{N_f,\gamma}(\mathbf{y}) \rangle_{\tilde{\mathcal{H}}^{(m)}_{N_f}} \\ &= \frac{1}{N_f} \sum_{l \in [m]} \sum_{q \in N_f} \sum_{\substack{\mathbf{i} \in \Delta_l(n_x-1) \\ \mathbf{j} \in \Delta_l(n_y-1)}} \langle \Delta\tilde{x}^{1,q}_{\mathbf{i}_1} \otimes \cdots \otimes \Delta\tilde{x}^{l,q}_{\mathbf{i}_l}, \Delta\tilde{y}^{1,q}_{\mathbf{i}_1} \otimes \cdots \otimes \Delta\tilde{y}^{l,q}_{\mathbf{j}_l} \rangle_{(\mathbb{R}^2)^{\otimes l}} \end{aligned}$$

It shall be convenient to introduce some extra notation. Similar to the syntax in many programming languages, we now use square brackets to denote indexing of tensors. For example, if $A \in \mathbb{R}^{p \times p}$ and $1 \leq i \leq j \leq p$, then $A[i:j, i:j]$ refers to the $(j-i+1)\times(j-i+1)$ sub-matrix consisting of the rows and columns from the i'th to the j'th index. We may

also write $A[i:j]$ to refer to the $(j-i+1)\times p$ sub-matrix consisting of the *rows* from the i'th to the j'th index. When writing $[:]$ we simply include all indices, that is, in order to get the sub-matrix consisting of the *columns* from the i'th to the j'th index, we would write $A[:, i:j]$. Note that this also implies that $A[i:j] = A[i:j,:]$. Thus, when we do not specify a specific index selection for the trailing dimensions, it should be understood as including all indices. We may also choose to select only a single index of a dimension, in which case the result is a tensor of one less dimension. For example, $A[i]$ corresponds to the p-dimensional vector that is the i'th row of A and, similarly, $A[:, j]$ corresponds to the j'th column. Finally, sometimes when assigning new values to a sub-tensor, we shall use ellipsis to make it clear that the new values are tensors themselves. So, if $a \in \mathbb{R}^p$, we would write $A[i, \dots] \leftarrow a$ for the new matrix $\tilde{A} \in \mathbb{R}^{p\times p}$ resulting from replacing the i'th column of A with a.

3.D.1 Signature stream

Throughout, we fix some $\mathbf{x} = (x_1, \dots, x_n)$ and define $\tilde{\mathbf{x}}$ as in Definition 16 (given some sample of the weights and for arbitrary values of the hyperparameters N_f and γ). We shall discuss an algorithm for computing the tensors

$$\mathbf{S}^{(l)}[k, q, t] := \sum_{\mathbf{i}\in\Delta_l(t)} \Delta\tilde{x}^{k,q}_{\mathbf{i}_1} \otimes \cdots \otimes \Delta\tilde{x}^{k+l,q}_{\mathbf{i}_l}$$

for $l \in [m]$, $k \in [m-l]$, $q \in [N_f]$, and $t \in [n_x - m - 1]$. Note that in order to compute the RFSF-DP map we only require the terms corresponding to $k = 1$. However, as we shall see, there are two main reasons for computing and storing the other terms as well. The first reason is that it will allow us to calculate the signature stream in the reverse direction using the exact same algorithm up to a simple reversal of terms. This point will sound a little vague for now, but we elaborate on it in Remark 12. The second reason is that it will allow us to compute the RFSF-DP map over sliding windows of the sequence $\mathbf{x}$ at the same cost as it would take to compute a single application of the map to the entire sequence. This is essentially the content of Algorithm 4.

Returning, for now, to the tensors $\mathbf{S}^{(l)}[k, q, t]$, one quickly derives the recursive relation

$$\mathbf{S}^{(l)}[k, q, t+1] = \mathbf{S}^{(l)}[k, q, t] + \mathbf{S}^{(l-1)}[k, q, t] \otimes \Delta\tilde{x}^{k+l,q}_{t+1}. \tag{3.D.1}$$

Thus, defining $\mathbf{a}[k, q, t] := \mathbf{S}^{(l-1)}[k, q, t-1]$ for $2 \le t \le n-1$ and $\mathbf{a}[k, q, 1] := 0 \in (\mathbb{R}^2)^{\otimes(l-1)}$, we obtain the update rule in Algorithm 1 which computes $\mathbf{S}^{(l)}[k, \dots]$ given $\mathbf{a}$ and $\Delta\tilde{\mathbf{x}}$. Given this update rule it is then easy to construct an algorithm computing $\mathbf{S}^{(l)}$ for all $l \in [m]$ just taking care to treat the case $l = 1$ separately. Given these tensors, one can then find $\tilde{S}^{(m)}_{N_f,\gamma}(\mathbf{x}[1:t])$ where $\mathbf{x}[s:t] = (x_s, \dots, x_t)$ simply by taking $\mathbf{S}^{(l)}[1, \dots]/\sqrt{N_f}$ for all $l \in [m]$. See also Algorithm 2.

Remark 11. The initialization step of Algorithm 2 has a run-time complexity $O(nN_fmp)$. For each level $l \in [m]$, the update given in Algorithm 1 requires on the order of nN_f2^l

Algorithm 1 Update

Input: $\mathbf{a}$, $\Delta\tilde{\mathbf{x}}$
  for $1 \leq q \leq N_f$ **do**
    $\mathbf{Q}[q, \dots] \leftarrow 0 \in ((\mathbb{R}^2)^{\otimes l})^{n-1}$
    $\mathbf{Q}[q, 1, \dots] \leftarrow \mathbf{a}[q, 1] \otimes \Delta\tilde{\mathbf{x}}[q, 1]$
    for $2 \leq t \leq n-1$ **do**
      $\mathbf{Q}[q, t, \dots] \leftarrow \mathbf{Q}[q, t-1, \dots] + \mathbf{a}[q, t] \otimes \Delta\tilde{\mathbf{x}}[q, t]$
    end for
    $\mathbf{a}'[q, 1, \dots] \leftarrow 0 \in (\mathbb{R}^2)^{\otimes l}$
    $\mathbf{a}'[q, 2 : (n-1), \dots] \leftarrow \mathbf{Q}[q, 1 : (n-2)] \in ((\mathbb{R}^2)^{\otimes l})^{n-2}$
  end for
Output: $\mathbf{Q}$, $\mathbf{a}'$

Algorithm 2 ForwardRFSF

Input: $\mathbf{x}$, m, γ, N_f, `flatten`

  # Initialize variables
  for $1 \leq l \leq m$ **do**
    for $1 \leq q \leq N_f$ **do**
      $w[l, q, \dots] \leftarrow \mathcal{N}(0, \gamma I_p)$
      $\tilde{\mathbf{x}}[l, q, \dots] \leftarrow (\cos(\mathbf{x}.w[l, q],\ \sin(\mathbf{x}.w[l, q]) \in (\mathbb{R}^2)^n$
      $\Delta\tilde{\mathbf{x}}[l, q, \dots] \leftarrow \tilde{\mathbf{x}}[l, q, 2 : n] - \tilde{\mathbf{x}}[l, q, 1 : (n-1)] \in (\mathbb{R}^2)^{n-1}$
      $\mathbf{a}[l, q, \dots] \leftarrow 1 \in \mathbb{R}^{n-1}$
    end for
  end for

  # Compute the RFSF-DP map
  for $1 \leq l \leq m$ **do**
    for $1 \leq k \leq m - l$ **do**
      $\mathbf{Q},\ \mathbf{a}' \leftarrow \text{Update}(\mathbf{a}[k], \Delta\tilde{\mathbf{x}}[l + k])$
      $\mathbf{S}^{(l)}[k, \dots] \leftarrow \mathbf{Q}/\sqrt{N_f}$
      $\mathbf{a}[k, \dots] \leftarrow \mathbf{a}'$
    end for
    if `flatten` **then**
      $\mathbf{S}^{(m)} \leftarrow \mathbf{S}^{(m)}[1]$
    end if
    $\mathbf{a} \leftarrow \mathbf{a}[1 : (m - l)]$
  end for

Output: $[\mathbf{S}^{(1)}, \dots,\ \mathbf{S}^{(M)}]$

computations. Thus, the two nested for-loops in the second block of Algorithm 2 requires on the order of

$$\sum_{l=1}^{m}\sum_{k=1}^{m-l} nN_f 2^l = O(nN_f 2^m)$$

computations. In total, we find that Algorith 2 has asymptotic run-time complexity $O(nN_f(mp + 2^m))$ which is exactly the same as Algorithm D.2 in Toth et al. [2023]. Note that it is linear in both the length of the sequence, n, and the dimension, p.

Remark 12. ForwardRFSF, as it is, computes the RFSF-DP map for $(x_1, \dots, x_t)$ for each $t \in [n]$. This is also sometimes referred to as the *signature stream* because the output can then be viewed as a stream in $\tilde{\mathcal{H}}_{N_f}^{(m)}$. One might only be interested in the signature over the full sequence $\mathbf{x} = (x_1, \dots, x_n)$, but the point is that in computing the full signature we get the whole stream for free. One might naturally ask if the stream in the other direction can be computed in the same way. That is, the stream whose t'th value corresponds to the RFSF-DP map of $(x_{n-t+1}, \dots, x_n)$. Upon little consideration, it is clear that the answer to this question must be yes. Indeed, let us define the tensors (note the subscript is $n - \mathbf{i}_j$ instead of $\mathbf{i}_j$)

$$\mathbf{S}_-^{(l)}[k, q, t] = \sum_{\mathbf{i} \in \Delta_l(t)} \Delta\tilde{x}_{n-\mathbf{i}_l}^{k,q} \otimes \cdots \otimes \Delta\tilde{x}_{n-\mathbf{i}_1}^{k+l,q}.$$

Then, similar to (3.D.1), one can derive the recursive relation

$$\mathbf{S}_-^{(l)}[k, q, t+1] = \mathbf{S}_-^{(l)}[k, q, t] + \Delta\tilde{x}_{n-t-1}^{k,q} \otimes \mathbf{S}_-^{(l-1)}[k-1, q, t]. \tag{3.D.2}$$

Comparing with (3.D.1), we note two differences. Firstly the tensor product on the right hand side has $\Delta\tilde{x}$ as the left term instead of the right term. The second difference is that we now require the $k-1$'th index of $\mathbf{S}^{(l-1)}$ for computing the k'th index of $\mathbf{S}^{(l)}$. As briefly mentioned earlier, this is one of the reasons why we chose to compute $\mathbf{S}^{(l)}[k, \dots]$ for all $k \leq m - l$. Now, based on (3.D.2), one may then define an update rule like Algorithm 1, but in the other direction, with only minor modifications. Furthermore, we can then construct an algorithm, call it BackwardRFSF, for computing the signature stream in the other direction completely analogous to ForwardRFSF.

3.D.2 First algorithm: Batched Signature MMD

We first present a very simple algorithm for computing (an approximate version of) $\hat{\varphi}_{sig}$. We shall call this algorithm *Batched Signature MMD* (or BaS-MMD). As presented here we assume that we are given a collection of an even number of windows $(\mathbf{x}_1, \dots, \mathbf{x}_{2B})$ all of which are of the same length $L \geq 1$. BaS-MMD then first splits up the batch into its two halves. On the first half it computes the average of the forward RFSF-DP stream (obtained from ForwardRFSF) and one the second half it computes the average of the backward RFSF-DP stream (obtained from BackwardRFSF). This yields two streams of length $L-1$ in $\tilde{\mathcal{H}}_{N_f}^{(m)}$. For each $t \in [L-1]$ we then compute the distance between the

values of the two streams at time t upon which we will have a vector of $L-1$ non-negative real numbers. One can then choose how to aggregate these numbers depending on the application. To obtain $\hat{\varphi}_{sig}$, for example, one would only compare the last values in the stream $t = L-1$. We leave the choice of aggregator open. That is, the BaS-MMD takes as its input some map $\text{agg} : \mathbb{R}_+^{L-1} \to \mathbb{R}_+$. One also has the option to first transform each window using some map $\xi : \mathbb{R}^p \to \mathbb{R}^{p'}$. This could, for example, be an ι-augmentation (see Def. 14). See Algorithm 3 for the specifics.

Both the run-time and the memory complexity are easily derived for BaS-MMD. Indeed, most of the computations for each window happen inside the signature algorithms ForwardRFSF and BackwardRFSF. Thus, referring to Remark 11, we see that the run-time complexity is $O(BLN_f(mp + 2^m))$.

3.D.3 Second algorithm: Sliding Window Signature MMD

The second algorithm we introduce is aptly named *Sliding Window Signature MMD* (or SWiS-MMD). It is a bit more involved than BaS-MMD, but also offers much greater flexibility in choosing the batching protocol. We shall no longer work on the level of signature streams, but now simply consider the RFSF-DP map over arbitrary windows of fixed length $L \geq 1$ over $\mathbf{x}$. Specifically, defining the tensors[15]

$$\mathbf{S}^{(l)}_{W,1}[k,q,t] := \sum_{\mathbf{i}\in\Delta_l(L)} \Delta\tilde{x}^{k,q}_{t+\mathbf{i}_1} \otimes \cdots \otimes \Delta\tilde{x}^{k+l,q}_{t+\mathbf{i}_l},$$
$$\mathbf{S}^{(l)}_{W,2}[k,q,t] := \sum_{\mathbf{i}\in\Delta_l(L)} \Delta\tilde{x}^{k,q}_{t+1+\mathbf{i}_1} \otimes \cdots \otimes \Delta\tilde{x}^{k+l,q}_{t+\mathbf{i}_l}$$

SWiS-MMD computes $\mathbf{S}^{(l)}_{W,1}[k,q,t]$ for all $l \in [m]$, $k \in [m-l]$, $q \in [N_f]$, and $t \in [n-L]$ and thus offers complete freedom in the choice of batching protocol. It relies on the key observation that

$$\mathbf{S}^{(l)}_{W,1}[k,q,t+1] = \mathbf{S}^{(l)}_{W,2}[k,q,t] + \mathbf{S}^{(l-1)}_{W,2}[k,q,t] \otimes \Delta\tilde{x}^{k+l,q}_{t+L+1}$$
$$\mathbf{S}^{(l)}_{W,2}[k,q,t+1] = \mathbf{S}^{(l)}_{W,1}[k,q,t+1] - \Delta\tilde{x}^{k,q}_{t} \otimes \mathbf{S}^{(l-1)}_{W,2}[k+1,q,t+1].$$

These two equations are exactly what gives the update rule in step 2 of Algorithm 4.[16] Starting with $\mathbf{S}^{(1)}_{W,1}[:,:,1] = \mathbf{S}^{(1)}_{-}[:,:,L]$ and $\mathbf{S}^{(1)}_{W,2}[:,:,1] = \mathbf{S}^{(2)}_{-}[:,:,L-1]$, we can then iterate over $2 \leq l \leq m$ and $2 \leq t \leq n-L$ to compute all the tensors. A subtle point here is that we actually use BacwkardRFSF to compute the signature over the initializing window since we need access to the RFSF-DP map over $(x_2, \ldots, x_L)$. The last detail of SWiS-MMD (or Algorithm 4) is the handling of arbitrary batching protocols. This can be done by specifying a batching map β. Assuming we have B_1 different batches each consisting of B_2 windows, the batching map, β, is a map from $[n-L]$ into $[B_1]$ so

[15] Note that the index t now specifies the starting point of the incoming slice of $\mathbf{x}$ and not the length as for $\mathbf{S}^{(l)}$.

[16] Note that, as with BackwardRFSF, computing the k'th index of the l'th level relies on having computed the $k+1$'th index of the $(l-1)$'th level.

Algorithm 3 BaS-MMD

Input: $(\mathbf{x}_1, \dots, \mathbf{x}_{2B})$, ξ, m, γ, N_f, agg

 # Step 1: Initialize variables
$L \leftarrow \text{length}(\mathbf{x}_1)$
for $1 \leq l \leq m$ **do**
  for $1 \leq t \leq L-1$ **do**
    $\Phi^{(l)}[t, \dots] \leftarrow 0 \in ((\mathbb{R}^2)^{\otimes l})^{N_f}$
    $\varphi[t] \leftarrow 0$
  end for
end for

 # Step 2: Compute average signatures over each half
for $1 \leq b \leq B$ **do**
  $\mathbf{u}, \mathbf{v} \leftarrow \xi(\mathbf{x}_b), \xi(\mathbf{x}_{B+b})$
  $[\mathbf{S}_+^{(1)}, \dots, \mathbf{S}_+^{(m)}] \leftarrow \text{ForwardRFSF}(\mathbf{u}, m, \gamma, N_f, \texttt{True})$
  $[\mathbf{S}_-^{(1)}, \dots, \mathbf{S}_-^{(m)}] \leftarrow \text{BackwardRFSF}(\mathbf{u}, m, \gamma, N_f, \texttt{True})$
  for $1 \leq l \leq m$ **do**
    for $1 \leq t \leq L-1$ **do**
      $\mathbf{s}_+ \leftarrow \mathbf{S}_+^{(l)}[:, t] / \|\mathbf{S}_+^{(l)}[:, t]\|$ ▷ Normalize signature
      $\mathbf{s}_- \leftarrow \mathbf{S}_-^{(l)}[:, t] / \|\mathbf{S}_-^{(l)}[:, t]\|$ ▷ Normalize signature
      $\Phi^{(l)}[t, \dots] \leftarrow \Phi^{(l)}[t] + (\mathbf{s}_+ - \mathbf{s}_-)/B$
    end for
  end for
end for

 # Step 3: Compute distance between average signatures and aggregate
for $1 \leq t \leq L-1$ **do**
  for $1 \leq l \leq m$ **do**
    $\varphi[t] \leftarrow \varphi[t] + \|\Phi^{(l)}[t]\|^2$
  end for
end for

Output: $\text{agg}(\varphi)$

that $\beta(t)$ specifies to which batch the window starting at t belongs to. We also allow β to return `None` in which case the window does not belong to any batch. In step 3, the average signatures over each batch are then computed and stored in a $B_1 \times B_1$ array and a user-specified aggregator is called on this matrix, e.g., a simple average.[17]

We find that the run-time complexity of SWiS-MMD is $O(nN_f(mp+2^m)+B_1^2 2^m)$, i.e., if the number of batches is not too large, comparable to computing a single call of the RFSF-DT map over the whole sequence and, in particular, linear in both sequence length and dimension. Were one to compute the RFSF-DT map over each window naively by calling ForwardRFSF (or BackwardRFSF) similar to BaS-MMD, the complexity would be $O(B_1 B_2 L N_f(mp + 2^m))$ which could potentially be much greater if there are a lot of overlapping windows.

3.E. Experiments

Below we present some finer details of the numerical experiments from Section 3.5. For more information on the exact implementation in `JAX`, we refer to (soon to be) public repository github.com/cholberg/stem where a notebook for each experiment can be found as well.

3.E.1 Example 3.5.1

The data was simulated as follows:

- For the diffusion matrix $\sigma = (\sigma_y, \sigma_z) \in \mathbb{R}^{3\times 3}$ we first randomly draw a 3 by 3 matrix with uniform entries between 0 and 1, call it $\tilde{\sigma}$. We then let $\sigma = \tilde{\sigma}\tilde{\sigma}^T + 2I$ where I is the 3-dimensional identity matrix.
- We then simulate a solution trajectory of (3.5.8) from $t_0 = 0$ to $T = 100$ using `diffrax` [Kidger, 2021] with a Euler solver of step size $\delta = 0.1$. Call (the linear interpolation of) these trajectories $\hat{y}_t$ and $\hat{z}_t$.
- Letting $\mathcal{T} = (10 = t_{burn} < \cdots < t_n = T)$ be a grid of $n = 2048$ equidistant points between 10 and 100, we then take $\mathbf{y} = (\hat{y}_t)_{t\in\mathcal{T}}$ and $\mathbf{z} = (\hat{z}_t)_{t\in\mathcal{T}}$. That is, we discard all points before a burn-in period.
- Finally, a sample $\mathbf{x}$ is obtained by employing the mixing d as described in Example 3.5.1 point-wise to the pair $(\mathbf{y}, \mathbf{z})$ with a radius $r = 10$.

To estimate the stationary embedding we first split the sequence $\mathbf{x}$ into 128 equal sized windows (of length $L = 2048/128 = 16$) $\mathbf{x}_1, \ldots, \mathbf{x}_{128}$ and then employ the BaS-MMD (see Algorithm 3) with ξ a path-wise min-max scaler, $m = 3$, $\gamma = .3$, $N_f = 64$, and agg being the simple average. Instead of calling BaS-MMD on the all the windows for each step of optimizer, we subsample a subset of size 64, call it $\mathbf{x}_{b_1}, \ldots, \mathbf{x}_{b_{64}}$, for each step.

[17] We are being slightly imprecise in that Algorithm 4 also assume that we have access to the weights w_q^l from BackwardRFSF since we are computing $\Delta\tilde{x}_t^{k,q}$ and $\Delta\tilde{x}_{t+L+1}^{k,q}$ for each window.

Algorithm 4 SWiS-MMD

Input: $\mathbf{x} = (x_1, \dots, x_n)$, L, β, B_1, B_2, ξ, m, γ, N_f, agg

 # Step 1: Initialize variables
$\mathbf{x}_\xi \leftarrow (\xi(x_i))_{1 \le i \le L}$
$[\mathbf{S}_-^{(1)}, \dots, \mathbf{S}_-^{(m)}] \leftarrow \text{BackwardRFSF}(\mathbf{x}_\xi, m, \gamma, N_f, \texttt{False})$
for $1 \le l \le m$ **do**
  $\mathbf{a}_1^{(l)} \leftarrow \mathbf{S}^{(l)}[:, :, L]$
  $\mathbf{a}_2^{(l)} \leftarrow \mathbf{S}^{(l)}[:, :, L-1]$
  $\mathbf{s} \leftarrow \mathbf{a}_1^{(l)}[1]$
  $\mathbf{s} \leftarrow \mathbf{s}/\|\mathbf{s}\|$ ▷ Normalize signature
  for $1 \le t \le B_1$ **do**
    $\Phi^{(l)}[t, \dots] \leftarrow 0 \in ((\mathbb{R}^2)^{\otimes l})^{N_f}$
  end for
  $b \leftarrow \beta(1)$
  if b is not `None` **then**
    $\Phi^{(l)}[b, \dots] \leftarrow \Phi^{(l)}[b] + \mathbf{s}/B_2$
  end if
end for

 # Step 2: Compute signatures over sliding window of length L
for $2 \le t \le n - L - m$ **do**
  $b \leftarrow \beta(t)$
  for $1 \le l \le m$ **do**
    for $1 \le q \le p_f$ **do**
      for $1 \le k \le m - l$ **do**
        $\mathbf{a}_1^{(l)}[k, q, \dots] \leftarrow \mathbf{a}_2^{(l)}[k, q] + \mathbf{a}_2^{(l-1)}[k, q] \otimes \Delta\tilde{\mathbf{x}}[l + k, q, t + L + 1]$
        $\mathbf{a}_2^{(l)}[k, q, \dots] \leftarrow \mathbf{a}_1^{(l)}[k, q] - \Delta\tilde{\mathbf{x}}[k, q, t] \otimes \mathbf{a}_2^{(l-1)}[k + 1, q]$
      end for
    end for
    $\mathbf{s} \leftarrow \mathbf{a}_1^{(l)}[1]$
    $\mathbf{s} \leftarrow \mathbf{s}/\|\mathbf{s}\|$ ▷ Normalize signature
    if b is not `None` **then**
      $\Phi^{(l)}[b, \dots] \leftarrow \Phi^{(l)}[b] + \mathbf{s}/B_2$
    end if
  end for
end for

 # Step 3: Compute distance between average signatures for each pair of batches and aggregate
for $1 \le b_1, b_2 \le B_1$ **do**
  for $1 \le l \le m$ **do**
    $\varphi[b_1, b_2] \leftarrow \|\Phi^{(l)}[b_1] - \Phi^{(l)}[b_2]\|^2$
  end for
end for

Output: $\text{agg}(\varphi)$

For the optimizer we use RMSprop with a learning rate of 0.001. We run the optimizer for 1000 epochs.

To compare trajectories of the estimated stationary component $\hat{y}$ with trajectories of the true latent process y_t in Figure 3.5.2, we shift and scale each path so that it starts at 0 and has a range of 1.

3.E.2 Example 3.5.2

The data was simulated as follows:

- For the diffusion matrix $\sigma = (\sigma_y, \sigma_z) \in \mathbb{R}^{3\times 3}$ we first randomly draw a 3 by 3 matrix with uniform entries between 0 and 1, call it $\tilde{\sigma}$. We then let $\sigma = (\tilde{\sigma}\tilde{\sigma}^T + 2I)/4$ where I is the 3-dimensional identity matrix.
- We then simulate a solution trajectory of (3.5.8) for each of the three choices $\theta = \theta_1, \theta_2, \theta_3$ from $t_0 = 0$ to $T = 100$ using `diffrax` [Kidger, 2021] with a Euler solver of step size $\delta = 0.1$. Call (the linear interpolation of) these trajectories $\hat{y}_t^i$ and $\hat{z}_t^i$ for $i = 1, 2, 3$.
- Letting $\mathcal{T} = (10 = t_{burn} < \cdots < t_n = T)$ be a grid of $n = 1024$ equidistant points between 10 and 100, we then take $\mathbf{y}^i = (\hat{y}_t^i)_{t\in\mathcal{T}}$ and $\mathbf{z}^i = (\hat{z}_t^i)_{t\in\mathcal{T}}$. That is, we discard all points before a burn-in period.
- We consider three random initializations of d^1, d^2, d^3. In each case, it is given by $d^j = d_1^j + d_2^j$ where $d_1^j : \mathbb{R} \to \mathbb{R}^3$ is a randomly intialized MLP with 1 hidden layer of width 256 and $d_2^j : \mathbb{R}^2 \to \mathbb{R}^3$ is defined similarly.
- Finally, we obtain the observations $\mathbf{x}^{i,j} = d^j(\mathbf{y}^i, \mathbf{z}^i)$ for all $i, j = 1, 2, 3$.

For each dataset $\mathbf{x}^{i,j}$ we fit the stationary embedding in the same way as in Example 3.5.1. In particular, we split our observation up into 64 blocks of length $L = 16$ obtaining $\mathbf{x}_1^{i,j}, \ldots, \mathbf{x}_{64}^{i,j}$. We then run stochastic gradient descent using the RMSprop optimizer from `optax` with a learning rate of 0.001 and a batch size of 32 for 1000 epochs. As above, to compare the estimated stationary processes to the true ones, we shift and normalize yielding Figure 3.5.4 as well as Figure 3.E.1 for the other two realizations of d.

3.E.3 Example 3.5.4

The data was simulated as follows:

- For the diffusion matrix $\sigma = (\sigma_y, \sigma_z) \in \mathbb{R}^{p\times p}$ we first randomly draw a p by p matrix with uniform entries between 0 and 1, call it $\tilde{\sigma}$. We then let $\sigma = \tilde{\sigma}\tilde{\sigma}^T/10 + 2I$ where I is the p-dimensional identity matrix.

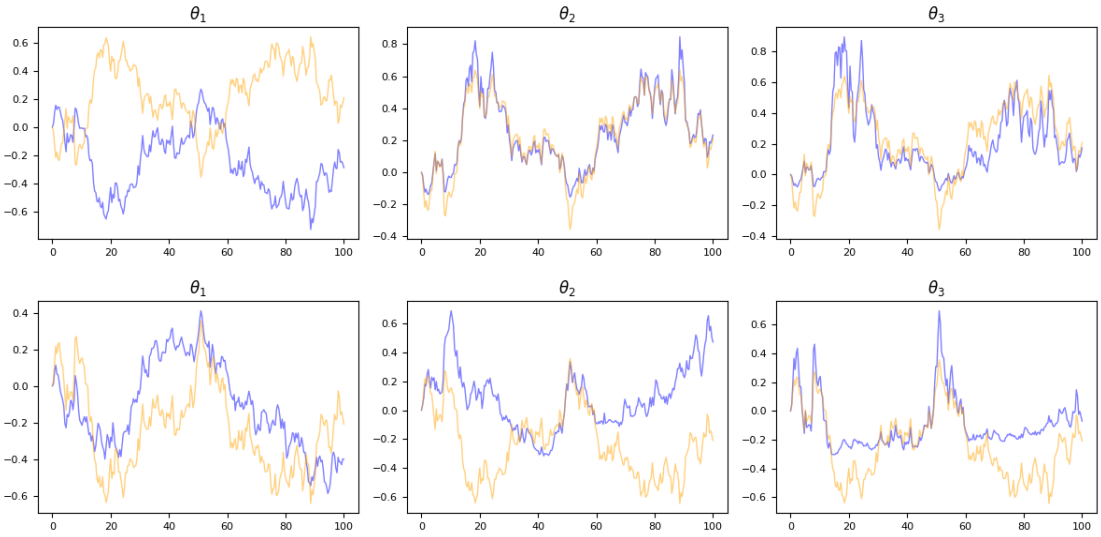

Figure 3.E.1: Estimating the stationary embedding of 3-dimensional non-stationary process. For each choice of θ, we plot a sample of the true stationary component y_t (orange) along with the corresponding estimate $\hat{y}_t = \hat{e}_1(x_t)$ (blue). Each row in the plot corresponds to a different realization of the mixing transformation d. All lines have been translated to start at 0 and normalized to have range 1.

- We then simulate a solution trajectory of (3.5.8) for each of the three choices

$$\begin{aligned}\theta_1(t,z) &= 0\\ \theta_2(t,z) &= 4\left(\sin\left(\frac{16\pi t}{T}\right),\ldots,\sin\left(\frac{16\pi tk}{T}\right)\right)\\ \theta_3(t,z) &= -z + t/4\end{aligned}$$

  from $t_0 = 0$ to $T = 64$ using `diffrax` [Kidger, 2021] with a Euler solver of step size $\delta = 0.1$. Call (the linear interpolation of) these trajectories $\hat{y}_t^i$ and $\hat{z}_t^i$ for $i = 1, 2, 3$. Similarly, we draw three trajectories from the p-dimensional OU-process $du_t = -u_t dt + dw_t^u$ using the same procedure resulting in $\hat{u}_t^i$ for $i = 1, 2, 3$. Here w_t^u is a p-dimensional standard BM independent of w_t in Eq. (3.5.8).

- Letting $\mathcal{T} = (10 = t_{burn} < \cdots < t_n = T)$ be a grid of $n = 1024$ equidistant points between 10 and 100, we then take $\mathbf{y}^i = (\hat{y}_t^i)_{t\in\mathcal{T}}$, $\mathbf{z}^i = (\hat{z}_t^i)_{t\in\mathcal{T}}$ and $\mathbf{u}^i = (\hat{u}_t^i)_{t\in\mathcal{T}}$. That is, we discard all points before burn-in period. We scale each of the three sequences so that all coordinates have range 1. We also shift $\mathbf{u}^i$ so that it starts at 0 and divide it by 2.

- We define the mixing transformation, $d : \mathbb{R}^p \to \mathbb{R}^p$, as a random initialization of a MLP with 1 hidden layer of width 128 and tanh activation function. Furthermore, we let g be given by $g(y) = \tanh(Wy+b)$ for $W \in \mathbb{R}^{p\times(p-k)}$ with elements uniformly

drawn between -0.5 and 0.5 and $b \in \mathbb{R}^p$ with elements uniformly drawn between 0 and 1.

- We let $\mathbf{x}^i = d(\mathbf{y}^i, \mathbf{z}^i)$ and $\mathbf{v}^i = g(\mathbf{y}^i) + \mathbf{u}^i$. We furthermore scale $\mathbf{x}^i$. We split up $\mathbf{v}^i$ into a training set, $\mathbf{v}^i_{train}$, consisting of the first 64 observations and a test set, $\mathbf{v}^i_{test}$, consisting of the last $1024 - 64$ observations.

For each of the three datasets $(\mathbf{v}^i_{train}, \mathbf{x}^i)$, we compare three different ways of estimating the regression function $f := g \circ e_1 : x \mapsto g(y)$. The first approach, the one we call `STEMd` (for STEM plus a decoder), first learns a stationary embedding $\hat{e}_1$, then computes $\hat{\mathbf{y}}^i = \hat{e}_1(\mathbf{x}^i)$, and finally estimates g by regressing $\mathbf{v}^i_{test}$ on the first 64 samples of $\mathbf{y}^i$. To learn the stationary embedding we split up $\mathbf{x}$ into 64 blocks as above. We then use the RMSprop optimizer form `optax` with a batch size of 32 and a learning rate of 1e-3 for the first 300 epochs and a learning rate of 1e-4 for the last 300 epochs. To ensure that the coordinates of $\hat{\mathbf{y}}^i$ are linearly independent, we add the penalty term $\text{pen} : \mathbb{R}^{L \times (p-k)} \to \mathbb{R}_+$ given by

$$\text{pen}(\mathbf{w}) = \frac{3/10^{-4}}{\det(\bar{\mathbf{w}}^\mathsf{T}\bar{\mathbf{w}})/\sqrt{L} + 10^{-4}}$$

where $\bar{\mathbf{w}}$ is the shifted and scaled version of $\mathbf{w}$ starting at 0 and with range 1. We estimate the embedding over the class of MLPs from p into $p - k$ with one hidden layer of size 128 and tanh activation functions. Finally, we scale and shift $\hat{\mathbf{y}}^i$ so that each coordinate has range 1 and starts at 0. For $W_{in} \in \mathbb{R}^{p \times (p-k)}$, $b_{in} \in \mathbb{R}^p$ and $W_{out} \in \mathbb{R}^{p \times p}$ and $b_{out} \in \mathbb{R}^p$, we define the function

$$g(y; W_{in}, b_{in}, W_{out}, b_{out}) = W_{out} \tanh\left(W_{in} y + b_{in}\right) + b_{out}.$$

We estimate the decoder, g, by $\hat{g} = g(\cdot; \hat{W}_{in}, \hat{b}_{in}, \hat{W}_{out}, \hat{b}_{out})$ where the parameters are found by regressing $\mathbf{v}^i_{train}$ on the first 64 values of $\hat{\mathbf{y}}^i$ using the RMSprop optimizer with a batch size of 32 and a learning rate of 0.01 for the first 500 epochs and a learning rate of 0.001 for the last 300 epochs to minimize the mean squared error. We also add an L^2-penalty term to the loss function penalizing W_{in} and b_{in}. This penalty term is weigthed by $\lambda = 0.1$.

For the other two approaches, we learn both the encoder, $\hat{e}_1$, and the decoder, $\hat{g}$, by regressing $\mathbf{v}^i_{train}$ on the first 64 values of $\mathbf{x}^i$. The procedure is exactly as for the first approach and the encoder is optimized over MLPs with one hidden layer of size 128 and tanh activation functions. For the approach which we denote `STEMae` (for auto-encoder with a stationary embedding), we add another penalty term (apart from the L^2-penalty) to our loss function which penalizes encoders that are not stationary. In particular, for each step of the optimizer we draw a batch of size 32 from $(\mathbf{v}^i_{train}, \mathbf{x}^i_{train})$ and also a batch of size 32 from $(\mathbf{x}^i_1, \ldots, \mathbf{x}^i_{64})$. The penalty term then evaluates BaS-MMD on the second batch. We weight this penalty term by $\lambda_{stat} = 0.01$. Finally, for the approach which we denote `autoencoder`, we simply fit the auto-encoder without the extra penalty term. In particular this approach does not see all the data in $\mathbf{x}^i_{test}$. We evaluate the three approaches by using the estimates to predict $\mathbf{v}^i_{test}$ on $\mathbf{x}^i_{test}$ and measuring the mean squared error. The results are reported in Table 3.5.1.

4 Beyond continuity: Differential equations with events

This chapter contains the paper:

[**STEM**] [Holberg and Salvi, 2024]. C. Holberg and C. Salvi. Exact gradients for stochastic spiking neural networks driven by rough signals. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

This is the first and only paper of the thesis not concerned with non-stationarity. Instead, it deals with the problem of deriving pathwise gradients of differential equation with *events* (as shall be defined shortly). Of particular interest is the application to spiking neural networks. Motivated by this goal, we define the *Marcus signature kernel maximum mean discrepancy*, a general way to compare two samples of càdlàg paths.

This paper, then, presents a way to deal with another form of irregularity often present in data: discontinuities.

Exact Gradients for Stochastic Spiking Neural Networks Driven by Rough Signals

Christian Holberg, Cristopher Salvi

Abstract

We introduce a mathematically rigorous framework based on rough path theory to model stochastic spiking neural networks (SSNNs) as stochastic differential equations with event discontinuities (Event SDEs) and driven by càdlàg rough paths. Our formalism is general enough to allow for potential jumps to be present both in the solution trajectories as well as in the driving noise. We then identify a set of sufficient conditions ensuring the existence of pathwise gradients of solution trajectories and event times with respect to the network's parameters and show how these gradients satisfy a recursive relation. Furthermore, we introduce a general-purpose loss function defined by means of a new class of signature kernels indexed on càdlàg rough paths and use it to train SSNNs as generative models. We provide an end-to-end autodifferentiable solver for Event SDEs and make its implementation available as part of the `diffrax` library. Our framework is, to our knowledge, the first enabling gradient-based training of SSNNs with noise affecting both the spike timing and the network's dynamics.

4.1 Introduction

Stochastic differential equations exhibiting event discontinuities (Event SDEs) and driven by noise processes with jumps are an important modelling tool in many areas of science. One of the most notable examples of such systems is that of stochastic spiking neural networks (SSNNs). Several models for neuronal dynamics have been proposed in the computational neuroscience literature with the *stochastic leaky integrate-and-fire* (SLIF) model being among the most popular choices [Gerstner and Kistler, 2002, Wunderlich and Pehle, 2021]. In its simplest form, given some continuous input current i_t on $[0, T]$, the dynamics of a single SLIF neuron consist of an Ornstein-Uhlenbeck process describing the membrane potential as well as a threshold for spike triggering and a resetting mechanism [Lansky and Ditlevsen, 2008]. In particular, between spikes, the dynamics of the membrane potential v_t is given by the following SDE

$$dv_t = \mu\,(i_t - v_t)\,dt + \sigma dB_t, \tag{4.1.1}$$

where $\mu > 0$ is a parameter and B_t is a standard Brownian motion. The neuron spikes whenever the membrane potential v hits the threshold $\psi > 0$ upon which v is reset to 0. Alternatively, one can model the spike times as a Poisson process with intensity $\lambda : \mathbb{R} \to \mathbb{R}_+$ depending on the membrane potential v_t. A common choice is $\lambda(v) = \exp((v - \psi)/\beta)$ [Pfister et al., 2006, Jimenez Rezende and Gerstner, 2014, Kajino, 2021, Jang and Simeone, 2022].

A notorious issue for calibrating Event SDEs such as SSNNs is that the implicitly defined event discontinuities, e.g., the spikes, make it difficult to define derivatives of the solution trajectories and of the event times with respect to the network's parameters using classical calculus rules. This issue is exacerbated when the dynamics are stochastic in which case the usual argument relying on the implicit function theorem, used for instance in Chen et al. [2020], Jia and Benson [2019], is no longer valid.

4.1.1 Contributions

In this paper, we introduce a mathematically rigorous framework to model SSNNs as SDEs with event discontinuities and driven by càdlàg rough paths, without any prior knowledge of the timing of events. The mathematical formalism we adopt is that of *rough path theory* Lyons [1998], a modern branch of stochastic analysis providing a robust solution theory for stochastic dynamical systems driven by noisy, possibly discontinuous, *rough signals*. Although Brownian motion is a prototypical example, these signals can be far more irregular (or rougher) than semimartingales Friz and Victoir [2010], Friz and Hairer [2020], Lyons et al. [2007].

Equipped with this formalism, we proceed to identify sufficient conditions under which the solution trajectories and the event times are differentiable with respect to the network's parameters and obtain a recursive relation for the exact pathwise gradients in Theorem 2. This is a strict generalization of the results presented in Chen et al. [2020] and Jia and Benson [2019] which only deal with ordinary differential equations (ODEs). Furthermore, we define *Marcus signature kernels* as extensions of continuous signature kernels Salvi et al. [2021a] to càdlàg rough paths and show their characteristicness. We then make use of this class of kernels indexed on discontinuous trajectories to define a general-purpose loss function enabling the training of SSNNs as generative models. We provide an end-to-end autodifferentiable solver for Event SDEs (Algorithm 1) and make its implementation available as part of the `diffrax` library [Kidger, 2021].

Our framework is, to our knowledge, the first allowing for gradient-based training of a large class of SSNNs where a noise process can be present in both the spike timing and the network's dynamics. In addition, we believe this work is the first enabling the computation of exact gradients for classical SNNs whose solutions are approximated via a numerical solver (not necessarily based on a Euler scheme). In fact, previous solutions are based either on surrogate gradients [Neftci et al., 2019] or follow an optimise-then-discretise approach deriving adjoint equations [Wunderlich and Pehle, 2021], the latter yielding exact gradients only in the scenario where solutions are available in closed form and not approximated via a numerical solver. Finally, we discuss how our results lead to bioplausible learning algorithms akin to *e-prop* [Bellec et al., 2020].

4.2 Related work

Neural stochastic differential equations (NSDEs) The intersection between differential equations and deep learning has become a topic of great interest in recent years. A neural ordinary differential equation (NODE) is an ODE of the form $\mathrm{d}y_t = f_\theta(y_t)dt$ started at $y_0 \in \mathbb{R}^e$ using a parametric Lipschitz vector field $f_\theta : \mathbb{R}^e \to \mathbb{R}^e$, usually given by a neural network Chen et al. [2018]. Similarly, a neural stochastic differential equation (NSDE) is an SDE of the form $\mathrm{d}y_t = \mu_\theta(y_t)dt + \sigma_\theta(y_t)dB_t$ driven by a d-dimensional Brownian motion B, started at $y_0 \in \mathbb{R}^e$, and with parametric vector field $\mu_\theta : \mathbb{R}^e \to \mathbb{R}^e$ and $\sigma_\theta : \mathbb{R}^e \to \mathbb{R}^{e\times d}$ that are Lip^1 and $\mathrm{Lip}^{2+\epsilon}$ continuous respectively[1]. Rough path theory offers a way of treating ODEs, SDEs, and more generally differential equations driven by signals or arbitrary (ir)regularity, under the unified framework of *rough differential equations* (RDEs) Morrill et al. [2021], Höglund et al. [2023]. For an account on applications of rough path theory to machine learning see Cass and Salvi [2024], Fermanian et al. [2023], Salvi [2021].

Training techniques for NSDEs Training a NSDE amounts to minimising over model parameters an appropriate notion of statistical divergence between a distribution of continuous trajectories generated by the NSDE and an empirical distribution of observed sample paths. Several approaches have been proposed in the literature, differing mostly in the choice of discriminating divergence. SDE-GANs, introduced in Kidger et al. [2021], use the 1-Wasserstein distance to train a NSDE as a Wasserstein-GAN Arjovsky et al. [2017]. Latent SDEs Li et al. [2020] train a NSDE with respect to the KL divergence via variational inference and can be interpreted as variational autoencoders. In Issa et al. [2023a] the authors propose to train NSDEs non-adversarially using a class of maximum mean discrepancies (MMD) endowed with signature kernels Király and Oberhauser [2019], Salvi et al. [2021a]. Signature kernels are a class of characteristic kernels indexed on continuous paths that have received increased attention in recent years thanks to their efficiency for handling path-dependent problems [Lemercier et al., 2021, Salvi et al., 2021c, Cochrane et al., 2021, Salvi et al., 2021b, Cirone et al., 2023, Pannier and Salvi, 2024, Manten et al., 2024]. For a treatment of this topic we refer the interested reader to [Cass and Salvi, 2024, Chapter 2]. These kernels are not applicable to sample trajectories of SSNNs because of the lack of continuity.

Backpropagation through NSDEs Once a choice of discriminator has been made, training NSDEs amounts to perform backpropagation through the SDE solver. There are several ways to do this. The first option is simply to backpropagate through the solver's internal operations. This method is known as *discretise-then-optimise*; it is generally speaking fast to evaluate and produces accurate gradients, but it is memory-inefficient, as every internal operation of the solver must be recorded. A second approach, known as *optimise-then-discretise*, computes gradients by deriving a backwards-in-time differential equation, the *adjoint equation*, which is then solved numerically by another call to the

[1] These are standard regularity conditions to ensure existence and uniqueness of a strong solution.

solver. Not storing intermediate quantities during the forward pass enables model training at a memory cost that is constant in depth. Nonetheless, this approach produces less accurate gradients and is usually slower to evaluate because it requires recalculating the forward solutions to perform the backward pass. A third way of backpropagating through NDEs is given by *algebraically reversible solvers*, offering both memory and accuracy efficiency. We refer to Kidger [2021] for further details.

Differential equations with events Many systems are not adequately modelled by continuous differential equations because they experience jump discontinuities triggered by the internal state of the system. Examples include a bouncing ball or spiking neurons. Such systems are often referred to as *(stochastic) hybrid systems* [Henzinger, 1996, Lygeros and Prandini, 2010]. When the differential equation is an ODE, there is a rich literature on sensitivity analysis aimed at computing derivatives using the implicit function theorem [Corner et al., 2019, 2020]. If, additionally, the vector fields describing the hybrid system are neural networks, Chen et al. [2020] show that NODEs solved up until first event time can be implemented as autodifferentiable blocks and Jia and Benson [2019] derive the corresponding adjoint equations. Nonetheless, none of these works cover the more general setting of SDEs. The only work, we are familiar with, dealing with sensitivity analysis in this setting is Pakniyat and Caines [2016], although focused on the problem of optimal control.

Training techniques for SNNs Roughly speaking, these works can be divided into two strands. The first, usually referred to as *backpropagation through time* (BPTT), starts with a Euler approximation of the SNN and does backpropagation by unrolling the computational graph over time; it then uses surrogate gradients as smooth approximations of the gradients of the non-differentiable terms. [Zenke and Vogels, 2021, Neftci et al., 2019, Ma et al., 2023]. This approach is essentially analogous to discretise-then-optimise where the backward pass uses custom gradients for the non-differentiable terms. The second strand computes exact gradients of the spike times using the implicit function theorem. These results are equivalent to optimise-then-discretise and can be used to define adjoint equations as in Wunderlich and Pehle [2021] or to derive forward sensitivities Lee et al. [2023]. However, we note that, unless solution trajectories and spike times of the SNN are computed exactly, neither method provides the actual gradients of the implemented solver. Furthermore, the BPTT surrogate gradient approach only covers the Euler approximation whereas many auto-differentiable differential equation solvers are available nowadays, e.g. in `diffrax`. Finally, there is a lot of interest in developing bioplausible learning algorithms where weights can be updated locally and in an online fashion. Notable advances in this direction include [Bellec et al., 2020, Xiao et al., 2022]. To the best of our knowledge, none of these works cover the case of stochastic SNNs where the neuronal dynamics are modeled as SDEs instead of ODEs.

4.3 Stochastic spiking neural networks as Event SDEs

We shall in this paper be concerned with SDEs where solution trajectories experience jumps triggered by implicitly defined events, dubbed *Event SDEs*. The prototypical example that we come back to throughout is the SNN model composed of SLIF neurons. Here the randomness appears both in the inter-spike dynamics as well as in the firing mechanism. To motivate the general definitions and concepts we start with an informal introduction of SSNNs.

4.3.1 Stochastic spiking neural networks

To achieve a more bioplausible model of neuronal behaviour, one can extend the simple deterministic LIF model by adding two types of noise: a diffusion term in the differential equation describing inter-spike behaviour [Lansky and Ditlevsen, 2008] and stochastic firing [Pfister et al., 2006, Kajino, 2021]. That is, the potential is modelled by eq. (4.1.1). Instead of firing exactly when the membrane potential hits a set threshold, we model the spike times (event times) by an inhomogenous Poisson process with intensity $\lambda : \mathbb{R}^e \to \mathbb{R}_+$ which is assumed to be bounded by some constant $C > 0$. This can be phrased as an Event SDE (note that this is essentially the reparameterisation trick) by introducing the additional state variable s_t satisfying

$$ds_t = \lambda(v_{t-})dt, \quad s_0 = \log u$$

where $u \sim \text{Unif}(0, 1)$. The neuron spikes whenever s_t hits 0 from below at which point the membrane potential is reset to a resting level and we sample a new initial condition for s_t. We can denote this first spike time by τ_1 and repeat the procedure to generate a sequence of spike times $\tau_1 < \tau_2 < ...$ In practice, we reinitialize s_t at $\log u - \alpha$ for some $\alpha > 0$. It can then be shown that

$$\mathbb{P}\left(t < \tau_{n+1} | \mathcal{F}_{\tau_n}\right) = \min\left\{1, \exp\left(\alpha - \int_{\tau_n}^{t} \lambda(v_{t-})dt\right)\right\} \quad \text{for } t \in [\tau_n, \tau_{n+1}).$$

It follows that $\tau_{n+1} - \tau_n \geq \alpha/C$ a.s., i.e. α controls the refractory period after spikes, a large value indicating a long resting period.

We can then build a SSNN by connecting such SLIF neurons in a network. In particular, apart from the membrane potential, we now also model the input current of each neuron as affected by the other neurons in the network. Let $K \geq 1$ denote the total number of neurons. We model neuron $k \in [K]$ be the three dimensional vector $y^k = (v^k, i^k, s^k)$ the dynamic of which in between spikes is given by

$$dv_t^k = \mu_1\left(i_t^k - v_t^k\right)dt + \sigma_1 dB_t^k, \quad di_t^k = -\mu_2 i_t^k dt + \sigma_2 dB_t^k, \quad ds_t^k = \lambda(v_t^k; \xi)dt, \quad (4.3.2)$$

where B^k is a standard two-dimensional Brownian motion, $\sigma = (\sigma_1, \sigma_2) \in \mathbb{R}^{2\times 2}$, $\mu = (\mu_1, \mu_2) \in \mathbb{R}^2$, and $\lambda(\cdot; \xi) : \mathbb{R} \to \mathbb{R}_+$ is an intensity function. As before, neuron k fires (or spikes) whenever s^k hits zero from below. Apart from resetting the membrane potential, this event also causes spikes to propagate through the network in a such a way that a

spike in neuron k will increment the input current of neuron j by w_{kj}. Here $w \in \mathbb{R}^{K\times K}$ is a matrix of weights representing the synaptic weights in the neural network. If one is only interested in specific network architectures such as, e.g., feed-forward, this can be achieved by fixing the appropriate entries in w at 0.

As presented here, there is no way to model information coming into the network. But this would only require a minor change. Indeed, by adding a suitable control term to eq. (4.3.2) we can model all relevant scenarios. Since this does not change the theory in any meaningful way (the general theory in Appendix 4.B covers RDEs so an extra smooth control is no issue), we only discuss the more simple model given without any additional input currents.

4.3.2 Model definition

Definition 1 (Event SDE)**.** Let $N \in \mathbb{N}$ be the number of events. Let $y_0 \in \mathbb{R}^e$ be an initial condition. Let $\mu : \mathbb{R}^e \to \mathbb{R}^e$ and $\sigma : \mathbb{R}^e \to \mathbb{R}^{e\times d}$ be the drift and diffusion vector fields. Let $\mathcal{E} : \mathbb{R}^e \to \mathbb{R}$ and $\mathcal{T} : \mathbb{R}^e \to \mathbb{R}^e$ be an event and transition function respectively. We say that $\left(y, (\tau_n)_{n=1}^N\right)$ is a solution to the Event SDE parameterised by $(y_0, \mu, \sigma, \mathcal{E}, \mathcal{T}, N)$ if $y_T = y_T^N$,

$$y_t = \sum_{n=0}^{N} y_t^n \mathbf{1}_{[\tau_n, \tau_{n+1})}(t), \quad \tau_n = \inf\left\{t > \tau_{n-1} : \mathcal{E}(y_t^{n-1}) = 0\right\}, \tag{4.3.3}$$

with $\mathcal{E}(y_{\tau_n}^n) \neq 0$ and

$$dy_t^0 = \mu(y_t^0)dt + \sigma(y_t^0)dB_t, \quad \text{started at } \; y_0^0 = y_0, \tag{4.3.4}$$

$$dy_t^n = \mu(y_t^n)dt + \sigma(y_t^n)dB_t, \quad \text{started at } \; y_{\tau_n}^n = \mathcal{T}\left(y_{\tau_n}^{n-1}\right), \tag{4.3.5}$$

where B_t is a d-dimensional Brownian motion and (4.3.4), (4.3.5) are Stratonovich SDEs.

In words, we initialize the system at y_0, evolve it using (4.3.4) until the first time τ_1 at which an event happens $\mathcal{E}(y_{\tau_1}^0) = 0$. We then transition the system according to $y_{\tau_1}^1 = \mathcal{T}\left(y_{\tau_1-}^0\right)$ and evolve it according to (4.3.5) until the next event is triggered. We note that Definition 1 can be generalised to multiple event and transition functions. Also, the transition function can be randomised by allowing it to have an extra argument $u \sim \mathrm{Unif}([0,1])$. As part of the definition we require that there are only finitely many events and that an event is not immediately triggered upon transitioning.

Existence of strong solutions to Event SDEs driven by continuous semimartingales has been studied in [Krystul and Blom, 2005, Theorem 5.2] and [Krystul et al., 2006]. Under sufficient regularity of μ and σ, a unique solution to (4.3.4) exists. We need the following additional assumptions:

Assumption 1. *There exists $c > 0$ such that for all $s \in (0,T)$ and $a \in im\,\mathcal{T}$ it holds that $\inf\{t > s \;:\; \mathcal{E}(y_t) = 0\} > c$ where y_t is the solution to 4.3.4 started at $y_s = a$*

Assumption 2. *It holds that $\mathcal{T}(ker\,\mathcal{E}) \cap \mathcal{E} = \emptyset$.*

Theorem 1 (Theorem 5.2, Krystul and Blom [2005]). *Under Assumptions 1-2 and with* $\mu \in \mathrm{Lip}^1$ *and* $\sigma \in \mathrm{Lip}^\gamma$ *for* $\gamma > 2$*, there exists a unique solution* $(y, (\tau_n)_{n=1}^N)$ *to the Event SDE of Definition 1.*

The definitions and results of this section can be extended to differential equations driven by random rough paths, and in particular, to cases where the driving noise exhibits jumps. In the latter case, it is important to note that the resulting Event SDE will exhibit two types of jumps: the ones given apriori by the driving noise and the ones that are implicitly defined through the solution (what we call *events*). In fact, we develop the main theory of Event RDEs in Appendix 4.A in the more general setting of RDEs driven by càdlàg rough paths. The rough path formalism enables a unified treatment of differential equations driven by noise signals of arbitrary (ir)regularity, and makes all proofs simple and systematic. In particular, it allows us to handle cases where the diffusion term is driven by a finite activity Lévy process (e.g, a homogeneous Poisson processes highly relevant in the context of SNNs).

4.3.3 Backpropagation

We are interested in optimizing a continuously differentiable loss function L whose input is the solution of a parameterised Event SDE. As for Neural ODEs, the vector fields, μ, σ, and the event and transition functions $\mathcal{E}, \mathcal{T}$, might depend on some learnable parameters θ. We can move the parameters θ of the Event RDE inside the initial condition y_0 by augmenting the dynamics with the additional state variable θ_t satisfying $d\theta_t = 0$ and $\theta_0 = \theta$. Thus, as long as we can compute gradients with respect to y_0, these will include gradients with respect to such parameters. We then require the gradients $\partial_{y_0} L$, if they exist. For this, we need to be able to compute the Jacobians $\partial y_t^n := \partial_{y_0} y_t^n$ of the inter-event flows associated to the dynamics of y_t^n and the derivatives $\partial \tau_n := \partial_{y_0} \tau_n$. We assume that the event and transition functions $\mathcal{E}$ and $\mathcal{T}$ are continuously differentiable.

Apriori, it is not clear under what conditions such quantities exist and even less how to compute them. This shall be the focus of the present section. We will need the following running assumptions.

Assumption 3. $\sigma(\mathcal{T}(y)) - \nabla\mathcal{T}(y)\sigma(y) = 0$ *for all* $y \in \ker \mathcal{E}$.

Assumption 4. $\nabla\mathcal{E}(y)\sigma(y) = 0$ *for all* $y \in \ker \mathcal{E}$.

Assumption 5. $\nabla\mathcal{E}(y)\mu(y) \neq 0$ *for all* $y \in \ker \mathcal{E}$.

Assumption 4 and 5 ensure that the event times are differentiable. Intuitively, they state that the event condition is hit only by the drift part of the solution. Assumption 4 holds for example if the event functions depend only on a smooth part of the system. Assumption 3 is what allows us to differentiate through the event transitions.

Theorem 2. *Let Assumptions 1-5 be satisfied and* $(y, (\tau_n)_{n=1}^N)$ *the solution to the Event SDE parameterized by* $(y_0, \mu, \sigma, \mathcal{E}, \mathcal{T}, N)$*. Then, almost surely, for any* $n \in [N]$*, the*

derivatives $\partial\tau_n$ *and the Jacobians* ∂y_t^n *exist and admit the following recursive expressions*

$$\partial\tau_n = -\frac{\nabla\mathcal{E}(y_{\tau_n}^{n-1})\partial y_{\tau_n}^{n-1}}{\nabla\mathcal{E}(y_{\tau_n}^{n-1})\mu(y_{\tau_n}^{n-1})} \tag{4.3.6}$$

$$\partial y_t^n = (\partial_{y_{\tau_n}^n} y_t^n)\left[\nabla\mathcal{T}(y_{\tau_n}^{n-1})\partial y_{\tau_n}^{n-1} - \left(\mu(y_{\tau_n}^n) - \nabla\mathcal{T}(y_{\tau_n}^{n-1})\mu(y_{\tau_n}^{n-1})\right)\partial\tau_n\right]. \tag{4.3.7}$$

where ∂y_t^n *and* $\partial\tau_n$ *are the total derivatives of* y_n^t *and* τ_n *with respect to the initial condition* y_0*,* $\partial_{y_{\tau_n}^n} y_t^n$ *denotes the partial derivative of the flow map of eq. (4.3.5) with respect to its intial condition, and* $\nabla\mathcal{T} \in \mathbb{R}^{e\times e}$ *and* $\nabla\mathcal{E} \in \mathbb{R}^{1\times e}$ *are the Jacobians of* $\mathcal{T}$ *and* $\mathcal{E}$*.*

Remark 1. If the diffusion term is absent we recover the gradients in Chen et al. [2020]. In this case, the assumptions of the theorem are trivially satisfied. Note however, that the result, as stated here, is slightly different since we are considering repeated events.

Remark 2. The recursive nature of (4.3.6) - (4.3.7) suggest a way to update gradients in an online fashion by computing the forward sensitivity along with the state of the Event SDE. In traditional machine learning applications (e.g. NDEs) forward mode automatic differentiation is usually avoided due to the fact that the output dimension tends to be orders of magnitude smaller than the number of parameters Kidger [2021]. However, for (S)SNNs this issue can be partly avoided as discussed in Section 4.4.4.

Returning now to the SSNN model introduced in Section 4.3.1 we find that it is an Event SDE with K different event functions given by $\mathcal{E}_k(y) = s^k$ and corresponding transition functions given by

$$\mathcal{T}_k(y) = \left(\mathcal{T}_k^1(y^1), \dots, \mathcal{T}_k^K(y^K)\right)$$

where $\mathcal{T}_k^j(y^j) = (v^j, i^j + w_{kj}, s^j)$ if $j \neq k$ and $\mathcal{T}_k^k(y^k) = (v^k - v_{reset}, i^k, \log u - \alpha)$ where $v_{reset} > 0$ is a constant determining by what amount the membrane potential is reset. The addition of the constant $\alpha > 0$ controlling the refractory period ensures that Assumption 2 and 2 are satisfied. Stochastic firing smoothes out the event triggering so that Assumption 5 and 4 hold. Finally, one can check that the combination of constant diffusion terms and the given transition functions satisfies Assumptions 3. Note that setting v_t^k exactly to 0 upon spiking would break Assumption 3. If one is interested in such a resetting mechanism it suffices to pick a diffusion term $\sigma_1(y^k)$ that satisfies $\sigma(0) = 0$. To sum up, solutions (in the sense of Def. 1) of the SSNNs exist and are unique. In addition, the trajectories and spike times are almost surely differentiable satisfying (4.3.6) and (4.3.7).

4.3.4 Numerical solvers

Theorem 2 gives an expression for the gradients of the event times as well as the Event SDE solution. In practice, analytical expressions for gradients are often not available and one has to resort to numerical solvers. Three solutions suggest themselves:

1. There are multiple autodifferentiable differential equation solvers (such as `diffrax` [Kidger, 2021]) that provide differentiable numerical approximations of the flows $\partial_{y^n_{\tau_n}} y^n_t$. We shall write $\text{SDESolve}(y_0, \mu, \sigma, s, t)$ for a generic choice of such a solver. Furthermore, if $\text{RootFind}(y_0, f)$ is a differentiable root finding algorithm (here $f : (y, t) \mapsto \mathbb{R}$ should be differentiable in both arguments and $\text{RootFind}(y_0, f)$ returns $t^* \in \mathbb{R}$ such that $f(y_0, t^*) = 0$), then we can define a differentiable map $E : y_0 \mapsto y^*$ by

$$t^* = \text{RootFind}\left(y_0, \mathcal{E}(\text{SDESolve}\left(\cdot, \mu, \sigma, s, \cdot\right))\right), \quad y^* = \text{SDESolve}\left(y_0, \mu, \sigma, s, t^*\right).$$

   Consequently, $\text{EventSDESolve}(y_0, \mu, \sigma, \mathcal{E}, \mathcal{T}, N)$ can be implemented as subsequent compositions of $\mathcal{T} \circ E$ (see Algorithm 1). This is a discretise-then-optimise approach [Kidger, 2021].

2. Alternatively, one can use the formulas (4.3.6) and (4.3.7) directly as a replacement of the derivatives. This is the approach taken in e.g. Chen et al. [2020]. To be precise, one would replace all the derivatives of the flow map (terms of the sort $\partial_{y^n_{\tau_n}} y^n_t$) with the derivatives of the given numerical solver. This approach is a solution between discretise-then-optimise and optimise-then-discretise.

3. Finally, one could apply the adjoint method (or optimise-then-discretise) as done for deterministic SNNs in Wunderlich and Pehle [2021] by deriving the adjoint equations. These adjoint equations define another SDE with jumps which is solved backwards in time. Between events the dynamics are exactly as in the continuous case so one just needs to specify the jumps of the adjoint process. This can be done by referring to (4.3.6) and (4.3.7).

Remark 3. One thing to be careful of with the discretise-then-optimise approach is that the SDE solver will compute time derivatives in the backward pass, although the modelled process is not time differentiable. Assumptions 4 and 3 should in principle guarantee that these derivatives cancel out (see Appendix 4.B), yet this might not necessarily happen at the level of the numerical solver because of precision issues. This is essentially due to the fact that approximate solutions provided by numerical solvers are in general not *flows*. Thus, when the path driving the diffusion term is very irregular, the gradients can become unstable. In practice we found this could be fixed by setting the gradient with respect to time of the driving Brownian motion to 0 and picking a step size sufficiently small.

Remark 4. In the context of SNNs, Algorithm 1 is actually a version of *exact backpropagation through time* (BPTT) of the unrolled numerical solution. Contrary to popular belief, this illustrates that one can compute exact gradients of numerical approximations of SNNs without the need to resort to surrogate gradient functions. Of course, this does not alleviate the so-called *dead neuron problem*. However, this ceases to be a problem when stochastic firing is introduced. In fact, surrogate gradients can be related to stochastic firing mechanisms and expected gradients Gygax and Zenke [2024].

Algorithm 1 EventSDESolve

Input $y_0, \mu, \sigma, \mathcal{E}, \mathcal{T}, N, t_0, \Delta t, T$

$y \leftarrow y_0$
$n \leftarrow 0$
$e \leftarrow \mathcal{E}(y)$
while $n < N$ **and** $t_0 < T$ **do**
  while $e < 0$ **do** ▷ We assume for simplicity that $e \leq 0$
    $y_0 \leftarrow y$
    $y \leftarrow \text{SDESolveStep}(y_0, \mu, \sigma, t_0, \Delta t)$
    $t_0 \leftarrow t_0 + \Delta t$
    $e \leftarrow \mathcal{E}(y)$ ▷ Update value of event function
  end while
  $t^*_{n+1} \leftarrow \text{RootFind}\left(y_0, \mathcal{E}(\text{SDESolveStep}\left(\cdot, \mu, \sigma, t_0 - \Delta t, \cdot\right))\right)$ ▷ Find exact event time
  $y^*_{n+1} \leftarrow \text{SDESolveStep}(y_0, \mu, \sigma, t_0 - \Delta_t, t^*_{n+1})$ ▷ Compute state at event time
  $y \leftarrow \mathcal{T}(y^*_{n+1})$ ▷ Apply transition function
  $n \leftarrow n + 1$
end while

Return $(t^*_n)_{n \leq N}$, y

4.4 Training stochastic spiking neural networks

4.4.1 A loss function based on signature kernels for càdlàg paths

To train SSNNs we will adopt a similar technique as in Issa et al. [2023a], where the authors propose to train NSDEs non-adversarially using a class of maximum mean discrepancies (MMD) endowed with signature kernels Salvi et al. [2021a] indexed on spaces of continuous paths as discriminators. However, as we mentioned in the introduction, classical signature kernels are not directly applicable to the setting of SSNNs as the solution trajectories not continuous. To remedy this issue, in Appendix 4.C, we generalise signature kernels to *Marcus signature kernels* indexed on discontinuous (or càdlàg) paths. We note that our numerical experiments only concern learning from spike trains, which are càdlàg paths of bounded variation. Yet, the Marcus signature kernel defined in Appendix 4.C can handle more general càdlàg rough paths.

The main idea goes as follows. If x is a càdlàg path, one can define the *Marcus signature* $S(x)$ in the spirit of Marcus SDEs [Marcus, 1978, 1981] as the signature of the *Marcus interpolation* of x. The general construction is given in Appendix 4.A. The *Marcus signature kernel* is defined as the inner product $k(x, y) = \langle S(x), S(y) \rangle$ of Marcus signatures $S(x), S(y)$ of two càdlàg paths x, y. As stated in the first part of Theorem 5, this kernel is characteristic on regular Borel measures supported on compact sets of càdlàg paths. In particular, this implies that the resulting *maximum mean discrepancy* (MMD)

$$d_k(\mu, \nu)^2 = \mathbb{E}_{x,x' \sim \mu} k(x, x') - 2\mathbb{E}_{x,y \sim \mu \times \nu} k(x, x') + \mathbb{E}_{y,y' \sim \nu} k(y, y')$$

satisfies the property $d_k(\mu,\nu)^2 = 0 \iff \mu = \nu$ for any two compactly supported measures μ, ν.

Nonetheless, characteristicness ceases to hold when one considers measures on càdlàg paths that are not compactly supported. In Chevyrev and Oberhauser [2022] the authors address this issue for continuous paths by using the so-called *robust signature*. They introduce a *tensor normalization* Λ ensuring that the range of the robust signature $\Lambda \circ S$ remains bounded. The *robust signature kernel* is then defined as the inner product $k_\Lambda(x,y) = \langle \Lambda \circ S(x), \Lambda \circ S(y) \rangle$. This normalization can be applied analogously to the *Marcus signature* resulting in a *robust Marcus signature kernel*. In the second part of Theorem 5, we prove characteristicness of k_Λ for possibly non-compactly supported Borel measures on càdlàg paths. The resulting MMD is denoted by d_{k_Λ}.

There are several ways of evaluating signature kernels. The most naive is to simply truncate the signatures at some finite level and then take their inner product. Another amounts to solve a path-dependent wave equation Salvi et al. [2021a]. Our experiments are compatible with both of these methods.

Given a collection of observed càdlàg trajectories $\{x^i\}_{i=1}^m \sim \mu^{\text{true}}$ sampled from an underlying unknown target measure μ^{true}, we can train an Event SDE by matching the generated càdlàg trajectories $\{y^i\}_{i=1}^n \sim \mu^\theta$ using an unbiased empirical estimator of d_k (or d_{k_Λ}), i.e. minimising over the parameters θ of the Event SDE the following loss function

$$\mathcal{L} = \frac{1}{m(m-1)} \sum_{j \neq i} k(x^i, x^j) - \frac{2}{mn} \sum_{i,j} k(x^i, y^j) + \frac{1}{n(n-1)} \sum_{j \neq i} k(y^i, y^j).$$

In the context of SSNNs, the observed and generated trajectories x^i's and y^i's correspond to spike trains, which are càdlàg paths of bounded variation.

4.4.2 Input current estimation

The first example is the simple problem of estimating the constant input current $c > 0$ based on a sample of spike trains in the single SLIF neuron model,

$$dv_t = \mu(c - v_t)dt + \sigma dB_t, \quad ds_t = \lambda(v_t)dt,$$

where $\lambda(v) = \exp(5(v-1)$, $\mu = 15$ and σ varies. Throughout we fix the true $c = 1.5$ and set $v_{reset} = 1.4$ and $\alpha = 0.03$. We run stochastic gradient descent for 1500 steps for two choices of the diffusion constant σ. The loss function is the signature kernel MMD between a simulated batch and the sample of spike trains.[2]. The test loss is the mean absolute error between the first three average spike times. Results are given in Fig. 4.4.1. For additional details regarding the experiments, we refer to Appendix 4.E.

In all cases backpropagation through Algorithm 1 is able to learn the underlying input current after around 600 steps up to a small estimation error. In particular, the convergence is fastest for the largest sample size and the true c is recovered for both levels of noise.

[2]For simplicity we only compute an approximation of the true MMD by truncating the signatures at depth 3 and taking the average across the batch/sample size.

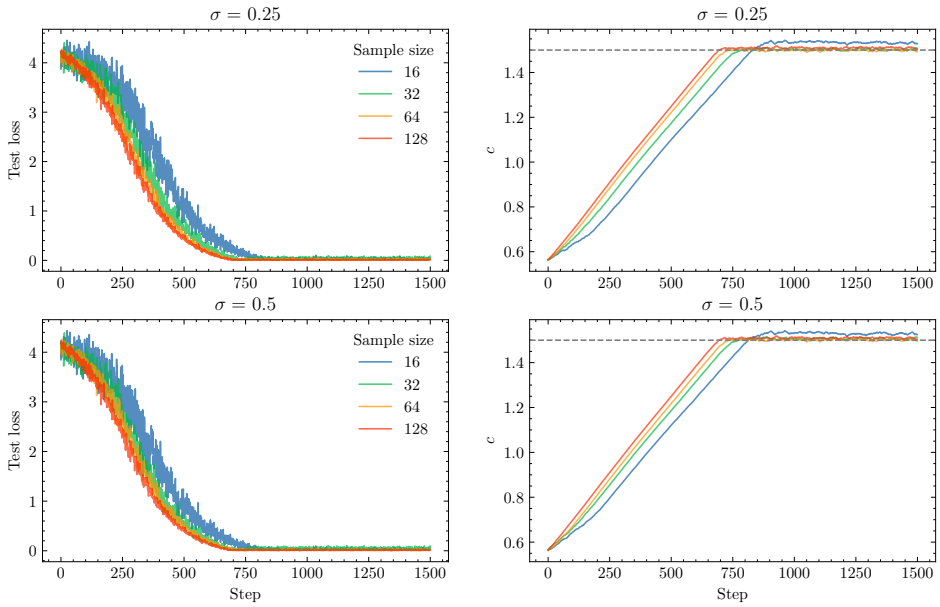

Figure 4.4.1: Test loss and c estimate across four sample sizes and for two levels of noise σ. On the left: MAE for the three first average spike times on a hold out test set. On the right: estimated value of c at the current step.

4.4.3 Synaptic weight estimation

Next we consider the problem of estimating the weight matrix in a feed-forward SSNN with input dimension 4, 1 hidden layer of dimension 16, and output dimension 2. The rest of the parameters are fixed throughout. We run stochastic gradient descent for 1500 steps with a batch size of 128 and for a sample size of $256, 512$, and 1024 respectively. Learning rate is decreased from 0.003 to 0.001 after 1000 steps. The results are given in Fig. 4.E.1 in Appendix 4.E. For a sample size of 512 and 1024 we are able to reach a test loss of practically 0, that is, samples from the learned model and the underlying model are more or less indistinguishable. Also, in all cases the estimated weight matrix approaches the true weight matrix. Interestingly, for the largest sample size, the model reaches the same test loss as the model trained on a sample size of 512, but their estimated weight matrices differ significantly.

4.4.4 Online learning

In the case of SSNNs, equations (4.3.6)-(4.3.7) lead to a formula for the forward sensitivity where any propagation of gradients between neurons only happens at spike times and only between connected neurons (see Proposition 1). Since the forward sensitivities are computed forward in time together with the solution of the SNN, gradients can be updated online as new data appears. As a result, between spikes of pre-synaptic neurons, we can update the gradient flow of the membrane potential and input current of each neuron using information exclusively from that neuron. For general network structures

and loss functions, however, this implies that each neuron needs to store on the order of K^2 gradient flows (one for each weight in the network).

On the other hand, if the adjacency matrix of the weight matrix forms a directed acyclic graph (DAG), three-factor Hebbian learning rules like those in Xiao et al. [2022], Bellec et al. [2020] are easily derived from Proposition 1. For simplicity, consider the SNN consisting of deterministic LIF neurons and let N_t^k denote the spike train of neuron k, i.e., N_t^k is càdlàg path equal to the number of spikes of neuron k at time t. We let $\tau^k(t)$ (or τ^k for short) denote the last spike of neuron k before time t. We shall assume that the instantaneous loss function L_t depends only on the most recent spike times $\tau^1, \dots, \tau^K$. Then,

$$\partial_{w_{jk}} L_t = \partial_{\tau^k} L_t \frac{a_{\tau^k}^{jk}}{\mu_1(v_{\tau^k}^k - i_{\tau^k}^k)}$$

where a_t^{jk} is the *eligibility trace* and the first term can be viewed as a *global modulator*, that is, a top-down learning signal propagting the error from the output neurons.[3] The eligibility trace satisfies

$$da_t^{jk} = \mu_1 \left(b_t^{jk} - a_t^{jk} \right) dt + \frac{v_{reset} a_t^{jk}}{\mu_1(i_t^k - v_t^k)} dN_t^k, \quad db_t^{jk} = -\mu_2 b_t^{jk} + dN_t^j,$$

where the dN terms are to be understood in the Riemann-Stieltjes sense. In other words, the eligibility trace can be updated exclusively from the activity of the pre and post-synaptic neurons. We note the similarity to the results derived in Bellec et al. [2020] only our result gives the exact gradients with no need to introduce surrogate gradient functions. For general network structures one can use the eligibility traces as proxies for the the true derivatives $\partial_{w_{ij}} \tau^k$.

4.5 Conclusion

We introduced a mathematical framework based on rough path theory to model SSNNs as SDEs exhibiting event discontinuities and driven by càdlàg rough paths. After identifying sufficient conditions for differentiability of solution trajectories and event times, we obtained a recursive relation for the pathwise gradients in Theorem 2, generalising the results presented in Chen et al. [2020] and Jia and Benson [2019] which only deal with the case of ODEs. Next, we introduced Marcus signature kernels as extensions of continuous signature kernels from Salvi et al. [2021a] to càdlàg rough paths and used them to define a general-purpose loss function on the space of càdlàg rough paths to train SSNNs where noise is present in both the spike timing and the network's dynamics. Based on these results, we also provided an end-to-end autodifferentiable solver for SDEs with event discontinuities (Algorithm 1) and made its implementation available as part of the `diffrax` repository. Finally, we discussed how our results lead to bioplausible

[3] Note that in the case of stochastic SNNs this term is not necessarily well-defined since semi-martingales are in general not differentiable wrt. time.

learning algorithms akin to *e-prop* [Bellec et al., 2020] but in the context of spike time gradients.

Despite the encouraging results we obtained, we think there are still many interesting research directions left to explore. For instance, we only made use of a discretise-then-optimise approach in our numerical experiments. It would be of interest to implement the adjoint equations or to use reversible solvers and compare the results. Similarly, since our Algorithm 1 differs from the usual approach with surrogate gradients even in the deterministic setting, questions remain on how these methods compare for training SNNs. Furthermore, it would be interesting to understand to what extent the inclusion of different types of driving noises in the dynamics of SSNNs would be beneficial for learning tasks compared to deterministic SNNs. Finally, it remains to be seen whether the discussion in Section 4.4.4 could lead to a bio-plausible learning algorithm with comparable performance to state-of-the-art backpropagation methods and implementable on neuromorphic hardware.

Appendix

The appendix is structured as follows. Section 4.A covers the basic concepts of càdlàg rough paths based on [Chevyrev and Friz, 2019] extended with a few of our own definitions and results. It culminates with the definition of Event RDEs which can be viewed as generalizations of Event SDEs. Section 4.B covers the proof of the main result, Theorem 2, but in the setting of Event RDEs as well as some preliminary technical lemmas needed for the proof. Section 4.C gives a brief overview of the main concepts in kernel learning and presents our results on Marcus signature kernels along with their proofs. Section 4.D derives the forward sensitivities of a SSNN. Finally, Section 4.E covers all the technical details of the simulation experiments that were not discussed in the main body of the paper.

4.A. Càdlàg rough paths

Marcus integration developed in Chevyrev and Friz [2019] preserves the chain rule and thus serves as an analog to Stratonovich integration for semi-martingales with jump discontinuities. In particular, it allows to define a canonical lift under which càdlàg semi-martingales are a.s. geometric rough paths and many of the results from the continuous case, such as universal limit theorems and stability results, carry over under suitably defined metrics.We briefly review some of the important concepts here by following the same setup as in Chevyrev and Friz [2019].

Let $C([0,T],E)$ and $D([0,T],E)$ be the space of continuous and càdlàg paths respectively on $[0,T]$ with values in a metric space (E,d). For $p \geq 1$, let $C_p([0,T],E)$ and $D_p([0,T],E)$ be the corresponding subspaces of paths with finite p-variation. For any $N \geq 1$, Let $G^N(\mathbb{R}^d)$ be the step-N free nilpotent Lie group over $\mathbb{R}^d$ endowed with the Carnot-Carathéodory metric d. Let $\Omega_p^C(\mathbb{R}^d) := C_p([0,T], G^{\lfloor p \rfloor}(\mathbb{R}^d))$ and $\Omega_p^D(\mathbb{R}^d) := D_p([0,T], G^{\lfloor p \rfloor}(\mathbb{R}^d))$ be the space of weakly geometric continuous and càdlàg p-rough paths respectively with the homogeneous p-variation metric

$$d_p(\mathbf{x},\mathbf{y}) = \max_{1\leq k\leq \lfloor p \rfloor} \sup_{\mathcal{D}\subset[0,T]} \left(\sum_{\mathcal{D}} d(\mathbf{x}_{t_i,t_{i+1}}, \mathbf{y}_{t_i,t_{i+1}})^{\frac{p}{k}} \right)^{\frac{k}{p}}.$$

Define the *log-linear* path function

$$\begin{aligned} \varphi : G^N(\mathbb{R}^d) \times G^N(\mathbb{R}^d) &\to C([0,1], G^N(\mathbb{R}^d)) \\ (\mathbf{a},\mathbf{b}) &\mapsto \exp((1-\cdot)\log \mathbf{a} + \cdot \log \mathbf{b}). \end{aligned}$$

where log and exp are the (truncated) tensor logarithm and exponential maps on $G^N(\mathbb{R}^d)$. If $N=1$, then $G^N(\mathbb{R}^d) \cong \mathbb{R}\oplus\mathbb{R}^d$ and $\varphi(a,b)_t = (1,(1-t)a+tb)$ is a straight line connecting a to b in unit time. For any $\mathbf{x} \in D([0,T],G^N(\mathbb{R}^d))$ we can construct a continuous path $\hat{\mathbf{x}} \in C([0,T],G^N(\mathbb{R}^d))$ by adding fictitious time and interpolating through the jumps using the log-linear path function according to the following definition.

Definition 2 (Marcus interpolation)**.** Let $N \geq 1$. For $\mathbf{x} \in D([0,T], G^N(\mathbb{R}^d))$, let $\tau_1, \tau_2, \dots, \tau_m$ be the jump times of $\mathbf{x}$ ordered such that $d(\mathbf{x}_{\tau_1-}, \mathbf{x}_{\tau_1}) \geq d(\mathbf{x}_{\tau_2-}, \mathbf{x}_{\tau_2}) \geq \cdots \geq d(\mathbf{x}_{\tau_m-}, \mathbf{x}_{\tau_m})$, where $0 \leq m \leq \infty$ is the number of jumps. Let (r_k) be a sequence of positive scalars $r_k > 0$ such that $r = \sum_{k=1}^m r_k < +\infty$. Define the discontinuous reparameterization $\eta : [0,T] \to [0, T+r]$ by

$$\eta(t) = t + \sum_{k=1}^{m} r_k \mathbf{1}_{\{\tau_k \leq t\}}.$$

The Marcus augmentation $\mathbf{x}^M \in C([0,T+r], G^N(\mathbb{R}^d))$ of $\mathbf{x}$ is the path

$$\mathbf{x}^M_s = \begin{cases} \mathbf{x}_t, & \text{if } s = \eta(t) \text{ for some } t \in [0,T], \\ \varphi(\mathbf{x}_{\tau_k-}, \mathbf{x}_{\tau_k})_{(s-\eta(\tau_k-))/r_k}, & \text{if } s \in [\eta(\tau_k-), \eta(\tau_k)) \text{ for } 1 \leq k < m+1. \end{cases}$$

The Marcus interpolation $\hat{\mathbf{x}} \in C([0,T], G^N(\mathbb{R}^d))$ of $\mathbf{x}$ is the path $\hat{\mathbf{x}} = \mathbf{x}^M \circ \eta_r$ where $\eta_r(t) = t(T+r)/T$ is a reparameterization from $[0,T]$ to $[0,T+r]$. We can recover $\mathbf{x}$ from $\hat{\mathbf{x}}$ via $\mathbf{x} = \hat{\mathbf{x}} \circ \eta_{\mathbf{x}}$ by considering the reparameterization $\eta_{\mathbf{x}} = \eta_r^{-1} \circ \eta$.

Once the Marcus interpolation is defined we can state what we mean by a solution to a differential equation driven by a geometric càdlàg rough path.

Definition 3 (Marcus RDE)**.** Let $\mathbf{x} \in \Omega_p^D(\mathbb{R}^d)$ and $f = (f_1, \dots, f_d)$ be Lip^γ vector fields on $\mathbb{R}^e$ with $\gamma > p$. For an initial condition $a \in \mathbb{R}^e$, let $\hat{y} \in C_p([0,T], \mathbb{R}^e)$ be the solution to the classical RDE driven by the Marcus interpolation $\hat{\mathbf{x}} \in \Omega_p^C(\mathbb{R}^d)$

$$d\hat{y}_t = f(\hat{y}_t) d\hat{\mathbf{x}}_t, \quad \hat{y}_0 = a.$$

Define the solution $y \in D_p([0,T], \mathbb{R}^e)$ to the Marcus RDE

$$dy_t = f(y_t) \diamond d\mathbf{x}_t, \quad y_0 = a \tag{4.A.1}$$

to be $y = \hat{y} \circ \eta_{\mathbf{x}}$, where $\eta_{\mathbf{x}}$ is the reparameterisation introduced in Definition 2.

4.A.1 Metrics on the space of càdlàg rough paths

Chevyrev and Friz [2019] introduce a metric α_p on $\Omega_p^D(\mathbb{R}^d)$ with respect to which 1) geometric càdlàg rough paths can be approximated with a sequence of continuous paths [Chevyrev and Friz, 2019, Section 3.2] and 2) the solution map $(y_0, \mathbf{x}) \mapsto (\mathbf{x}, y)$ of the Marcus RDE (4.A.1) is locally Lipschitz continuous [Chevyrev and Friz, 2019, Theorem 3.13].

We write Λ for the set of increasing bijections from $[0,T]$ to itself. For a $\lambda \in \Lambda$ we let $|\lambda| = \sup_{t \in [0,T]} |\lambda(t) - t|$. We first define the Skorokhod metric as well as a Skorokhod version of the usual p-variation metric.

Definition 4. For $p \geq 1$ and $\mathbf{x}, \mathbf{y} \in D_p([0,T], E)$, we define

$$\sigma_\infty(\mathbf{x}, \mathbf{y}) = \inf_{\lambda \in \Lambda} \max \left\{ |\lambda|, \sup_{t \in [0,T]} d((\mathbf{x} \circ \lambda)_t, \mathbf{y}_t) \right\},$$
$$\sigma_p(\mathbf{x}, \mathbf{y}) = \inf_{\lambda \in \Lambda} \max \left\{ |\lambda|, d_p\left(\mathbf{x} \circ \lambda, \mathbf{y}\right) \right\}.$$

It turns out that the topology induced by σ_p is too strong. In particular, it is not possible to approximate paths with jump discontinuities with a sequence of continuous paths (see Section 3.2 in Chevyrev and Friz [2019]). For $\mathbf{x} \in \Omega_p^D(\mathbb{R}^d)$ and $f = (f_1, \dots, f_d)$ a family of vector fields in $\mathrm{Lip}^{\gamma-1}(\mathbb{R}^e)$ with $\gamma > p$, let $\Phi_f(y, s, t; \mathbf{x})$ denote the solution to the Marcus RDE $dy_t = f(y_t) \diamond d\mathbf{x}_t$ initialized at $y_s = y$ and evaluated at time t. We define the set

$$J_f = \left\{ ((\mathbf{a}, b), (\mathbf{a}', b')) \mid \mathbf{a}, \mathbf{a}' \in G^{\lfloor p \rfloor}(\mathbb{R}^e), \Phi_f(b, 0, 1; \varphi(\mathbf{a}, \mathbf{a}')) = b' \right\}.$$

and, on it, the path function

$$\varphi_f((\mathbf{a}, b), (\mathbf{a}', b'))_t = \left(\varphi(\mathbf{a}, \mathbf{a}'), \Phi_f(b, 0, 1; \varphi(\mathbf{a}, \mathbf{a}')_t)\right).$$

Finally, we let $D_p^f([0,T], G^{\lfloor p \rfloor}(\mathbb{R}^d) \times \mathbb{R}^e)$ be the space of càdlàg paths $\mathbf{z} = (\mathbf{x}, y)$ on $G^{\lfloor p \rfloor}(\mathbb{R}^d) \times \mathbb{R}^e$ of bounded p-variation such that $(\mathbf{z}_t-, \mathbf{z}_t) \in J_f$ for all jump times t of $\mathbf{z}$. To keep notation simple, we shall write D_p^f when this does not cause any confusion. Naturally, if y is the solution to the Marcus RDE $dy_t = f(y_t) \diamond d\mathbf{x}_t$, we have $(\mathbf{x}, y) \in D_p^f$. For a $\mathbf{z} = (\mathbf{x}, y) \in D_p^f$ we may define the Marcus interpolation by interpolating the jumps using φ_f. Let $\hat{\mathbf{z}}^\delta$ denote this interpolation but with r_k replaced by δr_k for $\delta > 0$ and similarly for $\hat{\mathbf{x}}^\delta$ with $\mathbf{x} \in \Omega_p^D(\mathbb{R}^d)$.

Definition 5. For $f = (f_1, \dots, f_d)$ a family of vector fields in $\mathrm{Lip}^{\gamma-1}(\mathbb{R}^e)$ with $\gamma > p$, let $\mathbf{z}, \mathbf{z}' \in D_p^f$ with $\mathbf{z} = (\mathbf{x}, y)$ and $\mathbf{z}' = (\mathbf{x}', y')$ and define

$$\alpha_p(\mathbf{x}, \mathbf{x}') = \lim_{\delta \to 0} \sigma_p(\hat{\mathbf{x}}^\delta, \hat{\mathbf{x}}'^\delta),$$
$$\alpha_p(\mathbf{z}, \mathbf{z}') = \lim_{\delta \to 0} \sigma_p(\hat{\mathbf{z}}^\delta, \hat{\mathbf{z}}'^\delta).$$

Remark 5. It is proven in Chevyrev and Friz [2019] that in both cases the limit in α_p exists, is independent of the choice of r_k, and that it is indeed a metric on $\Omega_p^D(\mathbb{R}^d)$ resp. D_p^f.

Theorem 3 (Theorem 3.13 + Proposition 3.18, Chevyrev and Friz [2019])**.** *Let $f = (f_1, \dots, f_d)$ be a family of vector fields in $\mathrm{Lip}^{\gamma-1}(\mathbb{R}^e)$ with $\gamma > p$. Then,*

1. *The solution map*

$$\mathbb{R}^e \times (\Omega_p^D(\mathbb{R}^d), \alpha_p) \to (D_p^f, \alpha_p)$$
$$(y_0, \mathbf{x}) \mapsto \mathbf{z} = (\mathbf{x}, y)$$

   of the Marcus RDE $dy_t = f(y_t) \diamond d\mathbf{x}_t$ initialized at $y_0 \in \mathbb{R}^e$ is locally Lipschitz.

2. *On sets of bounded p-variation, the solution map*

$$\mathbb{R}^e \times (\Omega_p^D(\mathbb{R}^d), \sigma_\infty) \to (D_p([0,T], \mathbb{R}^e), \sigma_\infty)$$
$$(y_0, \mathbf{x}) \mapsto y$$

   of the Marcus RDE $dy_t = f(y_t) \diamond d\mathbf{x}_t$ initialized at $y_0 \in \mathbb{R}^e$ is continuous.

Now, let $C_0^1(\mathbb{R}^d)$ be the space of absolutely continuous functions on $\mathbb{R}^d$.

Definition 6. We define the space of geometric càdlàg p-rough paths $\Omega_{0,p}^D(\mathbb{R}^d)$ as the closure of $C_0^1(\mathbb{R}^d)$ in $\Omega_p^D(\mathbb{R}^d)$ under the metric α_p.

Remark 6. A càdlàg semi-martingale $x \in D_p([0,T],\mathbb{R}^d)$ can be canonically lifted to a geometric càdlàg p-rough path, with $p \in [2,3)$, by enhancing it with its two-fold iterated Marcus integrals, i.e.

$$\mathbf{x}_{s,t} = (1, x_{s,t}, \int_{s,t} (x_s - x_u) \otimes \diamond dx_u) \in G^2(\mathbb{R}^d)$$

where the integral is defined in a similar spirit to Definition 3 (see, for example, Friz and Zhang [2018] for more information). The solution to the corresponding Marcus RDE agrees a.s. with the solution to the usual càdlàg Marcus SDE which, in turn, if x has a.s. continuous sample paths, agrees a.s. with the solution to the Stratonovich SDE. See, e.g., Proposition 4.16 in Chevyrev and Friz [2019].

4.A.2 Signature

The extended tensor algebra over $\mathbb{R}^d$ is given by

$$T\left(\left(\mathbb{R}^d\right)\right) = \prod_{n=0}^{\infty} \left(\mathbb{R}^d\right)^{\otimes n}$$

equipped with the usual addition $+$ and tensor multiplication $\otimes$. An element $\mathbf{a} \in T\left(\left(\mathbb{R}^d\right)\right)$ is a formal series of tensors $\mathbf{a} = (\mathbf{a}^0, \mathbf{a}^1, \dots)$ such that $\mathbf{a}^n \in (\mathbb{R}^d)^{\otimes n}$. We define the projections $\pi_n : T\left(\left(\mathbb{R}^d\right)\right) \to (\mathbb{R}^d)^{\otimes n}$ given by $\pi_n(\mathbf{a}) = \mathbf{a}^n$. Let $\tilde{T}((\mathbb{R}^d))$ be the subset of $T((\mathbb{R}^d))$ such that the $\pi_0(\mathbf{a}) = 1$ for all $\mathbf{a} \in \tilde{T}((\mathbb{R}^d))$. Finally, we define the set of group-like elements,

$$G^{(*)} = \left\{ \mathbf{a} \in \tilde{T}\left(\left(\mathbb{R}^d\right)\right) \mid \pi_n(\mathbf{a}) \in G^N\left(\mathbb{R}^d\right) \text{ for all } n \geq 0 \right\}$$

Definition 7. Let $p \geq 1$ and $\mathbf{x} \in \Omega_p^D(\mathbb{R}^d)$. The signature of $\mathbf{x}$ is the path $S(\mathbf{x}) : [0,T] \mapsto G^{(*)}$ such that, for each $N \geq 0$,

$$dS(\mathbf{x})_t^N = S(\mathbf{x})_t^N \otimes \diamond d\mathbf{x}_t, \quad S(\mathbf{x})_0^N = \mathbf{1} \in G^N\left(\mathbb{R}^d\right). \tag{4.A.2}$$

Remark 7. Uniqueness and existence of the signature follow from the continuous analog. Indeed, by definition, (4.A.2) is equivalent to a continuous linear RDE.

Remark 8. The signature, as defined here, is also known as the *minimal jump extension of* $\mathbf{x}$ and was first introduced in Friz and Shekhar [2017]. It was further explored in Cuchiero et al. [2022] where it was also shown that it acts as a universal feature map.

In the continuous case, it is well known that the signature characterizes paths up to *tree-like equivalence*. Two continuous paths $\mathbf{x}, \mathbf{y}$ are said to be tree-like equivalent if there exists a continuous non-negative map $h : [0,T] \to \mathbb{R}_+$ such that $h(0) = h(T)$ and

$$\|\mathbf{x}_{s,t} - \mathbf{y}_{s,t}\| \leq h(s) + h(t) - 2 \inf_{u \in [s,t]} h(u).$$

This can be generalized to càdlàg paths in the following way. We say that two càdlàg paths $\mathbf{x}, \mathbf{y}$ are tree-like equivalent, or $\mathbf{x} \sim_t \mathbf{y}$, if their their Marcus interpolations (see Def. 2), $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$, are tree-like equivalent. It is straightforward to check that this indeed is an equivalence relation on $\Omega_p^D(\mathbb{R}^d)$. Perhaps more interestingly, we obtain the following result. For ease of notation we shall henceforth mean $S(\mathbf{x})_T$ when omitting the subscript from the singature.

Proposition 1. *Let $p \geq 1$. The map $S(\cdot) : \Omega_p^D(\mathbb{R}^d) \to G^{(*)}$ is injective up to tree-like equivalence, i.e., $S(\mathbf{x}) = S(\mathbf{y})$ iff $\mathbf{x} \sim_t \mathbf{y}$.*

Proof. The result follows from the continuous case upon realizing that $S(\mathbf{x}) = S(\hat{\mathbf{x}})$ and analogously for $\mathbf{y}$. □

4.A.3 Young pairing

In many cases, given a geometric càdlàg rough path $\mathbf{x} \in \Omega_{0,p}^D(\mathbb{R}^d)$ with $p \in [2,3)$ and a path $h \in D_1([0,T], \mathbb{R}^e)$ of bounded variation one is interested in constructing a new rough path $\mathbf{y} \in \Omega_{0,p}^D(\mathbb{R}^{d+e})$ such that the first level of $\mathbf{y}$ is given by $y = (x, h)$. In the continuous case this can be done by using the level two information $\mathbf{x}^2$ and $\int dh \otimes dh$ to fill in the corresponding terms in $\mathbf{y}^2$ and using the well-defined Young cross-integrals to fill in the rest. The resulting level 2 rough path is called the *Young pairing* of $\mathbf{x}$ and h and we will denote it by $\mathbf{y} = P(\mathbf{x}, h)$. The canonical example to keep in the mind is when $h_t = t$, that is, we want to augment the rough path with an added time coordinate (see Def. 9). In the càdlàg case one needs to be more careful in defining the appropriate Marcus lift.

Definition 8 (Definition 3.21 [Chevyrev and Friz, 2019])**.** Let $\mathbf{x} \in \Omega_{0,p}^D(\mathbb{R}^d)$ with $p \geq 1$ and $h \in D_1([0,T], \mathbb{R}^e)$. Define the path $\mathbf{z} = (\mathbf{x}, h)$ and the corresponding Marcus lift $\hat{\mathbf{z}} = (\hat{\mathbf{x}}, \hat{h})$. The Young pairing of $\mathbf{x}$ and h is the p-rough path $P(\mathbf{x}, h) \in \Omega_p^D(\mathbb{R}^{d+e})$ such that

$$P(\mathbf{x}, h) = P(\hat{\mathbf{x}}, \hat{h}) \circ \eta_{\mathbf{z}}$$

where $P(\hat{\mathbf{x}}, \hat{h})$ is the usual Young pairing of a continuous rough path and a continuous bounded variation path (see Def. 9.27 in Friz and Victoir [2010]).

We can then construct the time augmented rough path as the rough path obtained by the Young pairing with the simple continuous bounded variation path $h_t = t$. It turns out that this pairing is continuous as a map from $\Omega_{0,p}^D(\mathbb{R}^d)$ to $\Omega_{0,p}^D(\mathbb{R}^{d+1})$.

Definition 9. Let $\mathbf{x} \in \Omega_{0,p}^D(\mathbb{R}^d)$. The time augmented version of $\mathbf{x}$ is the unique rough path $\tilde{\mathbf{x}} \in \Omega_{0,p}^D(\mathbb{R}^{d+1})$ obtained by the Young pairing $P(\mathbf{x}, h)$ of $\mathbf{x}$ with the continuous bounded variation path $h_t = t$.

Proposition 2. *Let $p \in [1,3)$. Then, the map $\mathbf{x} \mapsto \tilde{\mathbf{x}}$ is continuous and injective as a map from $\Omega^D_{0,p}(\mathbb{R}^d)$ to $\Omega^D_{0,p}(\mathbb{R}^{d+1})$.*

Proof. Let $\mathcal{X} = \Omega^D_{0,p}(\mathbb{R}^d)$ be a metric space when equipped with α_p. Fix $\mathbf{x} \in \mathcal{X}$ and let x^n be a sequence of absolutely continuous paths converging in $\mathcal{X}$ to $\mathbf{x}$. We shall first show that $\tilde{x}^n$ then converges to $\tilde{\mathbf{x}}$. Since x^n does not have any jumps and any reparameterisation of x^n is still absolutely continuous, we may assume that

$$\alpha_p(\mathbf{x}, x^n) = \lim_{\delta \to 0} d_p(\hat{\mathbf{x}}^\delta, x^n) \to 0$$

for $n \to \infty$. Define $\mathbf{z} = (\mathbf{x}, h)$ and $\hat{\mathbf{z}}^d = (\hat{\mathbf{x}}^\delta, \hat{h}^\delta)$ the Marcus interpolation with $\eta_{\mathbf{x},\delta}$ the reparameterisation such that $\mathbf{z} = \hat{\mathbf{z}}^\delta \circ \eta_{\mathbf{x},\delta}$. Furthermore, let $P(\mathbf{x}, h)$ be the Young pairing of $\mathbf{x}$ and h. By definition,

$$\begin{aligned}
\alpha_p\left(P(\mathbf{x},h), P(x^n,h)\right) &= \lim_{\delta\to 0} \sigma_p\left(P(\hat{\mathbf{x}}^\delta, \hat{h}^\delta), P(x^n,h)\right) \\
&\leq \lim_{\delta\to 0} d_p\left(P(\hat{\mathbf{x}}^\delta, \hat{h}^\delta), P(x^n,h)\right) \\
&\leq \lim_{\delta\to 0} C\left(d_p(\hat{\mathbf{x}}^\delta, x^n) + d_1(\hat{h}^\delta, h)\right) \to 0
\end{aligned}$$

for $n \to \infty$ where C is just some generic constant depending only on p. The last inequality follows from 9.32 in Friz and Victoir [2010]. Thus, if $\mathbf{y} \in \mathcal{X}$ is such that $\alpha_p(\mathbf{x}, \mathbf{y}) < \epsilon$, we can choose another sequence y^n of absolutely continuous paths and $N \geq 1$ large enough so that

$$\alpha_p\left(P(\mathbf{x},h), P(\mathbf{y},h)\right) \leq 2\epsilon + \alpha_p(P(x^n,h), P(y^n,h)),$$

$$\alpha_p(x^n, y^n) \leq 2\epsilon$$

for all $n \geq N$. By Remark 3.6 in Chevyrev and Friz [2019], we then have that, up to choosing a large N, $d_p(x^n, y^n) \leq \epsilon$ for all $n \geq N$ and therefore, once more appealing to Theorem 9.32 in Friz and Victoir [2010],

$$\alpha_p\left(P(x^n,h), P(y^n,h)\right) \leq 2C\epsilon.$$

In conclusion, $\alpha_p\left(P(\mathbf{x},h), P(\mathbf{y},h)\right) \leq 2(1+C)\epsilon$ which proves the result.

Injectivity follows from Cuchiero et al. [2022]. □

4.A.4 Event RDEs

The results of Section 4.3.3 hold in more generality. In fact, we can define Event RDEs similar to Definition 1 where the inter-event dynamics are given by Marcus RDEs driven by càdlàg rough paths. Utilizing the correspondence between solutions to Marcus RDEs and Marcus SDEs, it then follows that the results in the main body of the paper are a special case of the results given below.

Definition 10 (Event RDE)**.** Let $p \geq 1$ and $N \in \mathbb{N}$ be the number of events. Let $\mathbf{x} \in \Omega_p^D(\mathbb{R}^d)$ and $f = (f_1, \dots, f_d)$ be a family of Lip^γ on $\mathbb{R}^e$ with $\gamma > p$. Let $\mathcal{E} : \mathbb{R}^e \to \mathbb{R}$ and $\mathcal{T} : \mathbb{R}^e \to \mathbb{R}^e$ be an event and transition function respectively. We say that $\left(y, (\tau_n)_{n=1}^N\right)$ is a solution to the Event RDE parameterised by $(y_0, \mathbf{x}, f, \mathcal{E}, \mathcal{T}, N)$ if $y_T = y_T^N$,

$$y_t = \sum_{n=0}^{N} y_t^n \mathbf{1}_{[\tau_n, \tau_{n+1})}(t), \quad \tau_n = \inf\left\{t > \tau_{n-1} : \mathcal{E}(y_{t-}^{n-1}) = 0\right\}, \tag{4.A.3}$$

with $\mathcal{E}(y_{\tau_n}^n) \neq 0$ and

$$dy_t^0 = f(y_t^0) \diamond d\mathbf{x}_t, \quad \text{started at} \quad y_0^0 = y_0, \tag{4.A.4}$$

$$dy_t^n = f(y_t^n) \diamond d\mathbf{x}_t, \quad \text{started at} \quad y_{\tau_n}^n = \mathcal{T}\left(y_{\tau_n -}^{n-1}\right). \tag{4.A.5}$$

Existence and uniqueness of solutions to Event RDEs is proven in the same way as for Event SDEs. Indeed, under the usual assumption that the vector fields f are Lip^γ, for $\gamma > p$, a unique solution to (4.A.4) exists. In fact, the solution map $y_s \times (s,t) \mapsto y_t$ is a diffeomorphism for every fixed $0 \leq s < t \leq T$ (see, e.g., Theorem 3.13 in Chevyrev and Friz [2019]). It follows that we can iteratively define a unique sequence of solutions $y^n \in D_p([t_n, T], \mathbb{R}^d)$. Finally, as mentioned in Remark 6, if the driving rough path $\mathbf{x}$ is the Marcus lift of a semi-martingale, the inter-event solutions agree almost surely with the solutions to the corresponding Marcus SDE.

Theorem 4. *Under Assumptions 1-2, there exists a unique solution $\left(y, (\tau_n)_{n=1}^N\right)$ to the Event RDE of Definition 10. Furthermore, if $\mathbf{x}$ is the Marcus lift of a Brownian motion, the solution coincides almost surely with the solution to the corresponding Event SDE as given in Def. 1.*

Hence, the Event SDEs considered in the main text are special cases of Event RDEs driven by the Marcus lift of a Brownian motion. Yet, the more general formulation of Event RDEs allows to treat, using the same mathematical machinery of rough path theory a much larger family of driving noises such as fractional Brownian motion or even smooth controls. Also, since the driving rough path is allowed to be càdlàg, the model class given by Def. 10 includes cases where the inter-event dynamics are given by Marcus SDEs driven by general semi-martingales.

4.B. Proof of Theorem 2

The proof of Theorem 2 presented below covers the case where $\left(y, (\tau_n)_{n=1}^N\right)$ is the solution to an Event RDE. Throughout we consider vector fields $\mu \in \text{Lip}^1, \sigma \in \text{Lip}^{2+\epsilon}$ and specialise to Event RDEs where the inter-event dynamics are given by

$$dy_t^n = \mu(y_t^n)dt + \sigma(y_t^n) \diamond d\mathbf{x}_t, \tag{4.B.1}$$

where $\mathbf{x} \in \Omega_p^D(\mathbb{R}^d)$. The notation above deserves some clarification. One can define the vector field $f = (\mu, \sigma)$ and the Young pairing $\tilde{\mathbf{x}}_t$ of $\mathbf{x}$ and $h_t = t$. Assuming $\mu \in \text{Lip}^{2+\epsilon}$

we can then view y_t^n as the unique solution to the Marcus RDE

$$dy_t^n = f(y_t^n) \diamond \tilde{\mathbf{x}}_t.$$

Alternatively, if one is not ready to impose the added regularity on the drift μ, one can view 4.B.1 as a RDE with drift as in Ch. 12 in Friz and Victoir [2010]. To accomodate this more general case where the path driving the diffusion term might be 1) càdlàg and 2) is not restricted to be the rough path lift of a semi-martingale, we shall need the following two additional assumptions:

Assumption 6. *For any $n \in [N]$, there exists a non-empty interval $I_n = (\tau_n - \delta_n, \tau_n + \delta_n)$ such that $\mathbf{x}$ is continuous over I_n. In other words, the càdlàg rough path $\mathbf{x}$, does not jump in small intervals around the event times (τ_n).*

Assumption 7. *For all $0 \leq n \leq N$ we define $s_n = \tau_n - \delta_n/2$ and $t_n = \tau_{n+1} + \delta_{n+1}/2$. It holds that $\mathbf{x} \in \Omega_{0,p}^D([s_n, t_n], \mathbb{R}^d)$, i.e., $\mathbf{x}$ is a geometric p-rough path on the intervals $[s_n, t_n]$.*

Remark 9. Note that Assumption 6 trivially holds if $\mathbf{x}$ is continuous. Otherwise, it is enough to assume, e.g., that $\mathbf{x}$ is the Marcus lift of a *finite activity* Lèvy process. Furthermore, by the properties of the metric α_p, if $\mathbf{x}$ is the canonical Marcus lift of a semi-martingale $x \in D_p([s,t], \mathbb{R}^{d-1})$, then there exists a sequence (x^m) of piece-wise linear paths $x^m \in C_0^1([0,T], \mathbb{R}^{d-1})$ such that

$$\alpha_{p,[s_n,t_n]}(x^m, \mathbf{x}) \to 0 \quad \text{as} \quad m \to \infty \quad \text{a.s.}$$

See, e.g. [Chevyrev and Friz, 2019, Example 4.21]. The setting of Section 4.3.3 is therefore a special case of the setting considered here and Theorem 2 follows from the proof below.

We shall need two technical lemmas for the proof of 2

Lemma 1. *Assume that Assumptions 1-5 and 6-7 are satisfied. Then, there exists an open ball $B_0 \subset O$ such that the following holds:*

1. *For all $a \in B_0$, $|\tau(a)| = N$.*
2. *For any $n \in [N]$, the maps*

$$B_0 \ni a \mapsto \left(\tau_n(a), y^{n-1}_{\tau_n(a)}(a)\right) \quad \text{are continuous.}$$

3. *For the sequence (x^m) as given in Assumption 7 and $(y^m, (\tau_n^m)_{n=1}^N)$ the corresponding Event RDE solution, for all $n \in [N]$, it holds that*

$$\lim_{m\to\infty} \sup_{a \in B_0} \left(|\tau_n^m(a) - \tau_n(a)| + \left| y^{m,n-1}_{\tau_n^m(a)}(a) - y^{n-1}_{\tau_n(a)}(a) \right| \right) = 0.$$

Proof. Recall that $\Phi(y, s, t; \mathbf{x})$ is the solution map or flow of the differential equation

$$dy_u = f(y_u) \diamond d\tilde{\mathbf{x}}_u, \quad y_s = y$$

evaluated at time t. The first step will be to prove continuity at y_0. In particular, let $y_0^m \in O$ approach y_0 for m going to infinity and denote the solutions to the corresponding Event RDEs by $\left(y^m, (\tau_n^m)_{n=1}^{N_m}\right)$. We claim that $\lim_{m\to\infty} N_m = N$ and

$$\lim_{m\to\infty} \tau_n^m = \tau_n, \quad \lim_{m\to\infty} y_{\tau_n^m}^{m,n-1} = y_{\tau_n}^n.$$

To see this, note that, by Theorem 3, there exists a sequence $\lambda_m \in \Lambda$ of continuous reparameterizations such that $|\lambda_m| \to 0$ and

$$\sup_{(s,t)\in\Delta_T} |\Phi(y_0, s, t; \mathbf{x}) - \Phi(y_0^m, \lambda^m(s), \lambda^m(t); \mathbf{x})| \to 0 \tag{4.B.2}$$

for $m \to \infty$. Note, furthermore, that $\Phi(y_0^m, s, t; \mathbf{x} \circ \lambda_m) = \Phi(y_0^m, \lambda_m(s), \lambda_m(t); \mathbf{x})$ for all $(s, t) \in \Delta_T$. We let $\left(\tilde{y}^m, (\tilde{\tau}_n^m)_{n=1}^{N_m}\right)$ be the solution to the Event RDE where $(y_0, \mathbf{x})$ is replaced by $(y_0^m, \mathbf{x} \circ \lambda_m)$. It suffices to prove that, for all $1 \le n \le N$,

$$\lim_{m\to\infty} \tilde{\tau}_n^m = \tau_n, \quad \lim_{m\to\infty} \tilde{y}_{\tilde{\tau}_n^m}^{m,n-1} = y_{\tau_n}^n. \tag{4.B.3}$$

Indeed, since $\tilde{\tau}_n^m = \lambda_m^{-1}(\tau_n^m)$ and $|\lambda_m| \to 0$, it then follows that $\tau_n^m \to \tau_n$ for $m \to \infty$. Furthermore, we have $\tilde{y}_{\tilde{\tau}_n^m}^{m,n-1} = y_{\tau_n^m}^{m,n-1}$.

We shall proof (4.B.3) using an inductive argument. We have that

$$\begin{aligned} y_t^0 &= \Phi(y_0, 0, t; \mathbf{x}), \quad \forall t \in [0, \tau_1], \\ \tilde{y}_t^{m,0} &= \Phi(y_0^m, 0, t; \mathbf{x} \circ \lambda_m), \quad \forall t \in [0, \tilde{\tau}_1^m]. \end{aligned}$$

Now fix some $0 < \epsilon < \delta_1$ where δ_1 is given in Assumption 6. Note that $|\mathcal{E}(y_t^0)| > 0$ for all $t \in [0, \tau_1 - \epsilon]$ and therefore, by (4.B.2), it follows that there exists an $m_0 \in \mathbb{N}$ such that, for all $m \ge m_0$,

$$\inf_{t\in[0,\tau_1-\epsilon]} |\mathcal{E}(\Phi(y_0^m, 0, t; \mathbf{x} \circ \lambda_m))| > 0$$

so that $\tilde{\tau}_1^m \ge \tau_1 - \epsilon$. Next, for some small $0 < \eta < \epsilon$, Assumption 4 and the Mean Value Theorem imply the existence of $a_\eta^+ = r_\eta^+ y_{\tau_1}^0 - (1 - r_\eta^+) y_{\tau_1+\eta}^0$, and $a_\eta^- = r_\eta^- y_{\tau_1-\eta}^0 + (1 - r_\eta^-) y_{\tau_1}^0$ with $r_\eta^+, r_\eta^- \in (0, 1)$ such that

$$\begin{aligned} \mathcal{E}\left(y_{\tau_1+\eta}^0\right) &= \mathcal{E}\left(y_{\tau_1}^0\right) + \nabla\mathcal{E}(a_\eta^+) \int_{\tau_1}^{\tau_1+\eta} \mu(y_s^0) dy_s, \\ \mathcal{E}\left(y_{\tau_1-\eta}^0\right) &= \mathcal{E}\left(y_{\tau_1}^0\right) - \nabla\mathcal{E}(a_\eta^-) \int_{\tau_1-\eta}^{\tau_1} \mu(y_s^0) dy_s, \end{aligned}$$

But then, by Assumption 5 and the fact that $\mathcal{E}(y_{\tau_1}^0) = 0$, for η small enough, $\mathcal{E}(y_{\tau_1+\eta}^0)$ and $\mathcal{E}(y_{\tau_1-\eta}^0)$ must lie on different sides of 0. Assumption 6 and eq. (4.B.2) then yield

the existence of a $m_1 \geq m_0$ such that $\tilde{\tau}_1^m \leq \tau_1 + \eta \leq \tau_1 + \epsilon$ and $\inf_{t\in[0,\tau_1+\eta]} |\mathcal{E}(\tilde{y}_t^{m,0})| > 0$ for all $m \geq m_1$. It follows that $\tilde{\tau}_1^m \to \tau_1$. Finally, note that

$$\left|\tilde{y}^{m,0}_{\tilde{\tau}_1^m} - y^0_{\tau_1}\right| \leq \left|\tilde{y}^{m,0}_{\tilde{\tau}_1^m} - y^0_{\tilde{\tau}_1^m}\right| + \left|y^0_{\tilde{\tau}_1^m} - y^0_{\tau_1}\right|.$$

Another application of (4.B.2) shows that the first term on the right hand side goes to 0 for $m \to \infty$ and second term vanishes by Assumption 6.

To prove the inductive step, assume that (4.B.3) holds for $i \leq n$. For all $t \in [\tau_n, \tau_{n+1}]$ it holds that

$$\tilde{y}_t^{m,n} = \Phi\left(\mathcal{T}\left(\tilde{y}^{m,n-1}_{\tilde{\tau}_n^m}\right), \tilde{\tau}_n^m, t; \mathbf{x}\circ\lambda_m\right), \quad y_t^n = \Phi\left(\mathcal{T}\left(y^{n-1}_{\tau_n}\right), \tau_n, t; \mathbf{x}\right)$$

and, since $\tilde{y}^{m,n-1}_{\tilde{\tau}_n^m} \to y^{n-1}_{\tau_n}$, $\tilde{\tau}_n^m \to \tau_n$, and $\mathcal{T}$ is continuous,

$$\lim_{m\to\infty} \sup_{t\in[\tau_n,T]} \left|\Phi\left(\mathcal{T}\left(\tilde{y}^{m,n-1}_{\tilde{\tau}_n^m}\right), \tilde{\tau}_n^m, t; \mathbf{x}\circ\lambda_m\right) - \Phi\left(\mathcal{T}\left(y^{n-1}_{\tau_n}\right), \tau_n, t; \mathbf{x}\right)\right| = 0$$

whence the same argument as above proves that (4.B.3) also holds for $n+1$. This completes the proof of the claim.

Now, by continuity at y_0, it follows that there exists some small $r > 0$ such that for all $a \in B_r(y_0)$ it holds that $|\tau(a)| = N$ and $\tau_n(a) \in (\tau_n - \delta_n/2, \tau_n + \delta_n/2)$ for all $n \in [N]$ where δ_n is as in Assumption 6. Furthermore, since Assumption 1-5 and 6-7 still hold for $a \in B_r(y_0)$, the same argument as above can be applied to show that $\tau_n(a)$ and $y^{n-1}_{\tau_n(a)}(a)$ are continuous at a. This proves parts 1 and 2.

To prove part 3 we employ a similar induction argument to the one above. First, note that, by Theorem 3, there exists a constant $C > 0$ not depending on x such that

$$\alpha_{p,[0,t_0]}\left(y^{m,0}(a), y^0(a)\right) \leq C\alpha_{p,[0,t_0]}\left(x^m, \mathbf{x}\right).$$

Since the latter term does not depend on y and goes to 0 for m going to infinity, we find that

$$\lim_{m\to\infty} \sup_{a\in B_r(y_0)} \alpha_{p,[0,t_1]}\left(y^{m,0}(a), y^0(a)\right) = 0. \tag{4.B.4}$$

Recall, $y^{\delta,0}(a)$ is the continuous path obtained by the Marcus interpolation with δr_k instead of r_k and similarly for $y^{m,\delta,0}(a)$. Note that $y^{m,\delta,0}(a) = y^{m,0}(a)$ by continuity. Letting $\tau_1^m(a)$ and $\tau_1^\delta(a)$ denote the first event time of $y^{m,0}(a)$ and $y^{\delta,0}(a)$ respectively, we have, for all $m \in \mathbb{N}$

$$\sup_{a\in B_r(x_0)} |\tau_1^m(a) - \tau_1(a)| \leq \sup_{a\in B_r(y_0)} \lim_{\delta\to 0} \left(\left|\tau_1^m(a) - \tau_1^\delta(a)\right| + \left|\tau_1^\delta(a) - \tau_1(a)\right|\right).$$

Now, let $B_0 = B_r(y_0)$. Since $\tau_1(a) \in (\tau_1 - \delta_1/2, \tau_1 + \delta_1/2)$ for all $a \in B_0$ and $\mathbf{x}$ is continuous over this interval, it follows that $\left|\tau_1^\delta(a) - \tau_1(a)\right|$ goes to 0 as $\delta \to 0$ for each $a \in B_0$. Furthermore, by definition of the metric α_p, eq. (4.B.4), and the fact that $y_0^{m,0}(a) = a = y_0^0(a)$, for each $a \in B_0$, a similar argument as the one employed in

the beginning of the proof then shows that $|\tau_1^m(a) - \tau_1^\delta(a)| \to 0$ as $\delta \to 0$ and, thus, $\lim_{m\to\infty} \sup_{a\in B_0} |\tau_1^m(a) - \tau_1(a)| = 0$. Finally, starting from the inequality

$$\left|y^{m,0}_{\tau_1^m(a)}(a) - y^0_{\tau_1(a)}(a)\right| \le \left|y^{m,0}_{\tau_1^m(a)}(x) - y^{\delta,0}_{\tau_1^\delta(a)}(a)\right| + \left|y^{\delta,0}_{\tau_1^\delta(a)}(a) - y^0_{\tau_1(a)}(a)\right|$$

and taking the limit as $\delta \to 0$ and then the supremum over $x \in B_0$ on both sides, we can argue in exactly the same way to show that part 3 holds for $n = 1$. We can then argue by induction, just as in the first part of the proof, to show that it holds for all subsequent event times as well. Thus, the set B_0 satisfies all the stated requirements. □

Lemma 2. *Let Assumption 6 hold and x^m be as in Assumption 7. Then, for all $n \in [N]$ and $p' > p$,*

$$\lim_{m\to\infty} d_{p',[s,t]}(x^m, \mathbf{x}) = 0, \quad \textit{for any} \quad \tau_n - \delta_n/2 \le s < t \le \tau_n + \delta_n/2.$$

Proof. Fix some $n \in [N]$, $p' > p$ and $\tau_n - \delta_n/2 \le s < t \le \tau_n + \delta_n/2$. Note that, for any continuous reparameterization $\lambda \in \Lambda$, $m \in \mathbb{N}$, and $\delta > 0$, it holds that

$$d_{p',[s,t]}(x^m, \mathbf{x}) \le d_{p',[s,t]}(x^m, x^m \circ \lambda) + d_{p',[s_n,t_n]}(x^m \circ \lambda, \hat{\mathbf{x}}^\delta) + d_{p',[s,t]}(\hat{\mathbf{x}}^\delta, \mathbf{x}),$$

where $\hat{\mathbf{x}}^\delta$ is the Marcus interpolation of $\mathbf{x}$ over the interval $[s_n, t_n]$. Taking the infimum over $\lambda \in \Lambda$ and the limit as $\delta \to 0$ on both sides, we obtain

$$d_{p',[s,t]}(x^m, \mathbf{x}) \le \alpha_{p',[s_n,t_n]}(x^m, \mathbf{x}) + \lim_{\delta\to 0} d_{p',[s,t]}(\hat{\mathbf{x}}^\delta, \mathbf{x}).$$

The first term on the right hand side goes to 0 as $m \to \infty$ by Assumption 7. Furthermore, since, by Assumption 6, $\mathbf{x}$ is continuous on $(\tau_n - \delta_n, \tau_n + \delta_n)$, it follows that $d_{\infty,[s,t]}(\hat{\mathbf{x}}^\delta, \mathbf{x})$ goes to 0 for $\delta \to \infty$. But the result then follows from Proposition 8.15 and Lemma 8.16 in Friz and Victoir [2010]. □

Proof of Theorem 2. Step 1: Assume that $x \in C_1([0,T], \mathbb{R}^{d-1})$. By [Friz and Victoir, 2010, Theorem 4.4], the Jacobian ∂y_t^0 exists and satisfies (4.3.7) for all $t \in [0, \tau_1)$. We shall prove that relations (4.3.6) and (4.3.7) hold for all $n \in [N]$ by induction. Thus, assume that ∂y_t^k and $\partial \tau_k$ exist for all $t \in [\tau_k, \tau_{k+1})$ and $k \le n-1$ and satisfy the stated relations. To emphasise the dependence on the initial condition, we will sometimes use the notation $y^n = y^n(y_0)$ and $\tau_n = \tau_n(y_0)$ for the solution of the Event RDE started at y_0. We want to show that, for arbitrary $h \in \mathbb{R}^e$, the following limits

$$\lim_{\epsilon\to 0} \frac{\tau_n^\epsilon - \tau_n}{\epsilon} \quad \text{and} \quad \lim_{\epsilon\to 0} \frac{y_t^{n,\epsilon} - y_t^n}{\epsilon} \quad \text{for } t \in [\tau_n, \tau_{n+1})$$

exist and satisfy the stated expressions, where $\tau_n^\epsilon = \tau_n(y_0 + h\epsilon)$ and $y^{n,\epsilon} = y^n(y_0 + h\epsilon)$.

For any $\epsilon > 0$, because $\mathcal{E}$ is continuously differentiable, the Mean Value Theorem implies that there exists $c_\epsilon \in \mathbb{R}^e$ on the line connecting $y^{n-1}_{\tau_n}$ to $y^{n-1}_{\tau_n^\epsilon}$ and another

$c'_\epsilon \in \mathbb{R}^e$ on the line connecting $y^{n-1,\epsilon}_{\tau^\epsilon_n}$ to $y^{n-1}_{\tau^\epsilon_n}$ such that

$$\begin{aligned}
\mathcal{E}\left(y^{n-1}_{\tau_n}\right) &= \mathcal{E}\left(y^{n-1}_{\tau^\epsilon_n}\right) + \nabla\mathcal{E}(c_\epsilon)\left(y^{n-1}_{\tau_n} - y^{n-1}_{\tau^\epsilon_n}\right) \\
&= \mathcal{E}\left(y^{n-1}_{\tau^\epsilon_n}\right) + \nabla\mathcal{E}\left(c_\epsilon\right)\left(\mu(y^{n-1}_{\tau_n})(\tau_n - \tau^\epsilon_n) + \sigma\left(y^{n-1}_{\tau_n}\right)\left(x_{\tau_n} - x_{\tau^\epsilon_n}\right) + o(|\tau_n - \tau^\epsilon_n|)\right), \\
\mathcal{E}\left(y^{n-1,\epsilon}_{\tau^\epsilon_n}\right) &= \mathcal{E}\left(y^{n-1}_{\tau^\epsilon_n}\right) + \nabla\mathcal{E}(c'_\epsilon)\left(y^{n-1,\epsilon}_{\tau^\epsilon_n} - y^{n-1}_{\tau^\epsilon_n}\right) \\
&= \mathcal{E}\left(y^{n-1}_{\tau^\epsilon_n}\right) + \nabla\mathcal{E}(c'_\epsilon)\left(\epsilon\left(\partial y^{n-1}_{\tau_n}\right)h + o(\epsilon)\right),
\end{aligned}$$

where the last equality follows from the induction hypothesis. We have $\mathcal{E}(y^{n-1}_{\tau_n}) = 0 = \mathcal{E}(y^{n-1,\epsilon}_{\tau^\epsilon_n})$. Thus, by rearranging, we find that

$$\begin{aligned}
\frac{\tau^\epsilon_n - \tau_n}{\epsilon} &= -\frac{\nabla\mathcal{E}(y^{n-1}_{\tau_n})\partial y^{n-1}_{\tau_n} h}{\nabla\mathcal{E}(y^{n-1}_{\tau_n})\left(\mu(y^{n-1}_{\tau_n}) + \sigma(y^{n-1}_{\tau_n})\frac{x_{\tau_n} - x_{\tau^\epsilon_n}}{\tau_n - \tau^\epsilon_n}\right)} + o(1) \\
&= -\frac{\nabla\mathcal{E}(y^{n-1}_{\tau_n})\partial y^{n-1}_{\tau_n} h}{\nabla\mathcal{E}(y^{n-1}_{\tau_n})\mu(y^{n-1}_{\tau_n})} + o(1)
\end{aligned}$$

where the second equality follows from Assumptions 4 and 5.

Assume for now that $\tau^\epsilon_n < \tau_n$. By another application of the Mean Value Theorem, there exists $c_\epsilon \in \mathbb{R}^e$ on the line connecting $y^{n-1}_{\tau_n}$ to $y^{n-1}_{\tau^\epsilon_n}$ such that

$$\begin{aligned}
y^{n,\epsilon}_{\tau_n} - y^n_{\tau_n} &= y^{n,\epsilon}_{\tau_n} - \mathcal{T}\left(y^{n-1}_{\tau_n}\right) \\
&= y^{n,\epsilon}_{\tau_n} - \mathcal{T}\left(y^{n-1}_{\tau^\epsilon_n}\right) - \nabla\mathcal{T}(c_\epsilon)(y^{n-1}_{\tau_n} - y^{n-1}_{\tau^\epsilon_n}) \\
&= y^{n,\epsilon}_{\tau^\epsilon_n} + \mu(y^{n,\epsilon}_{\tau_n})(\tau_n - \tau^\epsilon_n) + \sigma\left(y^{n,\epsilon}_{\tau_n}\right)\left(x_{\tau_n} - x_{\tau^\epsilon_n}\right) - \mathcal{T}\left(y^{n-1}_{\tau^\epsilon_n}\right) \\
&\quad - \nabla\mathcal{T}(c_\epsilon)\left(\mu(y^{n-1}_{\tau_n})(\tau_n - \tau^\epsilon_n) + \sigma\left(y^{n-1}_{\tau_n}\right)\left(x_{\tau_n} - x_{\tau^\epsilon_n}\right) + o(|\tau_n - \tau^\epsilon_n|)\right) \\
&= \mathcal{T}\left(y^{n-1,\epsilon}_{\tau^\epsilon_n}\right) - \mathcal{T}\left(y^{n-1}_{\tau^\epsilon_n}\right) + \left(\mu(y^{n,\epsilon}_{\tau_n}) - \nabla\mathcal{T}(c_\epsilon)\mu(y^{n-1}_{\tau_n})\right)(\tau_n - \tau^\epsilon_n) \\
&\quad + \left(\sigma(y^n_{\tau_n}) - \nabla\mathcal{T}(c_\epsilon)\sigma(y^{n-1}_{\tau_n})\right)\left(x_{\tau_n} - x_{\tau^\epsilon_n}\right) + o(|\tau_n - \tau^\epsilon_n|)
\end{aligned}$$

Therefore

$$\frac{y^{n,\epsilon}_{\tau_n} - y^n_{\tau_n}}{\epsilon} = \nabla\mathcal{T}(y^{n-1}_{\tau_n})\partial y^{n-1}_{\tau_n} h + \left(\mu(y^n_{\tau_n}) - \nabla\mathcal{T}(y^{n-1}_{\tau_n})\mu(y^{n-1}_{\tau_n})\right)\partial\tau_n h + o(1)$$

where we used Assumption 3, the chain rule and the existence of $\partial\tau_n$. Finally, for any $t \in (\tau_n, \tau_{n+1}]$, equation (4.3.7) follows from the fact that we can write $y^n_t = \Phi(y^n_s, s, t, x)$ for all $\tau_n \le s < t$. In particular, by the chain rule, we find that

$$\partial y^n_t = \left[\partial_{y^n_s}\Phi(y^n_s, s, t)\partial y^n_s\right]_{s=\tau_n} = \partial_{y^n_{\tau_n}} y^n_t \left[\partial y^n_s\right]_{s=\tau_n}.$$

Step 2: Consider now the general case of $\mathbf{x} \in \Omega_p(\mathbb{R}^e)$ and let $(y^m, (\tau^m_{n_m})^{N_m}_{n_m})$ denote the solution to the Event RDE where $\mathbf{x}$ is replaced by the piece-wise linear approximation

x^m. With $\partial y_t^{n,m}$ and $\partial \tau_n^m$ denoting the corresponding derivatives, we saw in the previous step that both exist and satisfy (4.3.6)-(4.3.7). We let R_t^n and ρ_n denote the right hand side of (4.3.7) and (4.3.6) respectively. This step consists of proving that, for $n \in [N]$ and $t \in (\tau_n, t_n)$,

$$\lim_{m\to\infty} \{|\tau_n^m - \tau_n| + |y_t^{m,n} - y_t^n|\} = 0 \tag{4.B.5}$$

and for some open ball B_0 around y_0

$$\lim_{m\to\infty} \sup_{a\in B_0} \{|\partial\tau_n^m(a) - \rho_n(a)| + \|\partial y_t^{m,n}(a) - R_t^n(a)\|\} = 0. \tag{4.B.6}$$

By Lemma 1 and continuity of $\mathcal{T}$ we have that $\mathcal{T}(y^{m,n-1}_{\tau_n^m})$ converges to $\mathcal{T}(y^{n-1}_{\tau_n})$ and τ_n^m converges to τ_n as $m \to +\infty$. Then, because $y_t^n = \Phi(\mathcal{T}(y^{n-1}_{\tau_n}), \tau_n, t; \mathbf{x})$ and $y_t^{m,n} = \Phi(\mathcal{T}(y^{m,n-1}_{\tau_n^m}), \tau_n^m, t; x^m)$, equation (4.B.5) follows from, Lemma 2 and Corollary 11.16 in Friz and Victoir [2010]. In fact, since B_0 was constructed in Lemma 1 in such a way that $\tau_n(a) < t_n$ for all $a \in B_0$ we also get that

$$\lim_{m\to\infty} \sup_{a\in B_0} \left\| \partial_{y^n_{\tau_n(a)}(a)} \Phi\left(y^n_{\tau_n(a)}(a), \tau_n(a), t; \mathbf{x}\right) - \partial_{y^{m,n}_{\tau_n^m(a)}(a)} \Phi\left(y^{m,n}_{\tau_n^m(a)}(a), \tau_n^m(a), t; x^m\right) \right\| = 0.$$

by the same corollary in Friz and Victoir [2010]. Thus, to prove (4.B.6), it suffices to show that, for all $n \in \{1, ..., N\}$,

$$\lim_{m\to\infty} \sup_{a\in B_0} \left\| \partial y^{m,n-1}_{\tau_n^m(a)}(a) - R^{n-1}_{\tau_n(a)}(a) \right\| = 0.$$

We shall prove it using another inductive argument starting with $n = 1$. In this case it suffices to show that

$$\lim_{m\to\infty} \sup_{a\in B_0} \|\partial_a \Phi(a, 0, \tau_1(a); \mathbf{x}) - \partial_a \Phi(a, 0, \tau_1^m(a); x^m)\| = 0.$$

By [Chevyrev and Friz, 2019, Theorem 3.3] we know that the above holds if $\tau_1^m(a)$ and $\tau_1(a)$ are replaced by $\tau_1 + \delta_1/2$. Now let Φ^{-1} be the reverse of the flow map Φ, that is,

$$\Phi^{-1}(a_1, s, t; \mathbf{x}) = a_0 \Leftrightarrow \Phi(a_0, s, t; \mathbf{x}) = a_1.$$

From Lemma 1 it follows that $y^0_{\tau_1(a)}(a) = \Phi^{-1}(y^0_{t_0}(a), \tau_1(a), t_0; \mathbf{x})$ and, for m large enough, $y^{m,0}_{\tau_1^m(a)}(a) = \Phi^{-1}(y^{m,0}_{t_0}(a), \tau_1^m(a), t_0; x^m)$. But the result then follows from Lemma 2 and [Friz and Victoir, 2010, Corollary 11.16]. To prove the inductive step, assume that (4.B.6) holds for all $i \le n-1$. Again, by inspecting (4.3.6) and (4.3.7) and using the inductive assumption, one finds that it is enough to show that

$$\lim_{m\to\infty} \sup_{a\in B_0} \left\| \partial_{y^{n-1}_{\tau_{n-1}}} \Phi\left(y^{n-1}_{\tau_{n-1}}, \tau_{n-1}, \tau_n; \mathbf{x}\right) - \partial_{y^{m,n-1}_{\tau_{n-1}^m}} \Phi\left(y^{m,n-1}_{\tau_{n-1}^m}, \tau^m_{n-1}, \tau^m_n; x^m\right) \right\| = 0,$$

where we suppressed the dependence on a for notational simplicity. This is done exactly as for y^0 and completes the proof of Step 2.

Step 3: The third and final step is to combine Step 1 and 2 to finish the proof. So far we have proven that 1) the theorem holds for continuous paths of bounded variation and 2) $(\tau_n^m, y_t^{m,n})$ converges to (τ_n, y_t^n) and $(\partial\tau_n^m(a), \partial y_t^{m,n}(a))$ converges uniformly to $(\rho_n(a), R_t^n(a))$ over $a \in B_0$ for all $t \in (\tau_n, t_n)$ and $n \in [N]$. From these results it immediately follows that $(\tau_n(a), y_t^n(a))$ is differentiable at $a = y_0$ with derivatives given by $(\rho_n(y_0), R_t(y_0))$ for all $t \in (\tau_n, t_n)$. What is left to show then, is that this also holds for all other t. But this follows immediately from the chain rule upon realizing that, for any $\tau_n < s < \tau_n + \delta_n/2 < t < \tau_{n+1}$,

$$y_t^n = \Phi(y_s^n, s, t; \mathbf{x}) \Rightarrow \partial y_t^n = \partial_{y_s^n}\Phi(y_s^n, s, t; \mathbf{x})R_s^n(y_0) = R_t^n(y_0).$$

□

4.C. Kernel methods

We give here a brief outline of some of the most central concepts related to kernel methods. For a more in-depth introduction we refer the reader to Muandet et al. [2017], Schölkopf and Smola [2002], Berlinet and Thomas-Agnan [2011]. Let $\mathcal{X}$ be a topological space. We shall in this paper only be concerned with positive definite kernels, that is, symmetric functions $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ for which the Gram matrix is positive definite. To such a kernel one may associate a feature map $\mathcal{X} \to \mathbb{R}^{\mathcal{X}}$ such that $x \mapsto k_x = k(x, \cdot)$. A reproducing kernel Hilbert space (RKHS) is a Hilbert space $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ such that the evaluation functionals, $ev_x : f \mapsto f(x)$, are bounded for each $x \in \mathcal{X}$. For all positive definite kernels there is a unique RKHS $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ such that $f(x) = \langle k_x, f\rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$ and $x \in \mathcal{X}$. This is also known as the *reproducing property*. Furthermore, with H denoting the linear span of $\{k_x \mid x \in \mathcal{X}\}$, it holds that $\bar{H} = \mathcal{H}$, i.e., H is dense in $\mathcal{H}$. Two important properties of kernels are *characteristicness* and *universality*.

Definition 11. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite kernel. Denote by H the linear span of $\{k_x \mid x \in \mathcal{X}\}$ and let $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ be a topological vector space containing H and such that the inclusion map $\iota : H \to \mathcal{F}$ is continuous.

- We say that k is universal to $\mathcal{F}$ if the embedding of $\iota : H \to \mathcal{F}$ is dense.
- We say that k is characteristic to $\mathcal{F}'$ if the embedding $\mu : \mathcal{F}' \to H'$, $D \mapsto D|_H$ is injective

Remark 10. This definition is the one used in Chevyrev and Oberhauser [2022] and is more general then the one usually encountered. Note that in many cases (all the cases considered here, in fact) $\mathcal{F}'$ will contain the set of probability measures on $\mathcal{X}$ in which case k being characteristic implies that the *kernel mean embedding* $\mu \mapsto \mathbb{E}_{X\sim\mu}k_X(\cdot)$ is injective.

Remark 11. Often times, instead of starting with the kernel function k and then obtaining the RKHS, one starts with a feature map $F : \mathcal{X} \to \mathcal{H}$ into a RKHS and then defines the kernel as the inner product in that Hilbert space, i.e., $k(x, y) = \langle F(x), F(y)\rangle_{\mathcal{H}}$. In

such cases, it makes sense to ask whether there are equivalent notions of F being universal and characteristic. This is indeed the case and the definition is almost the same as above. We refer to Definition 6 in Chevyrev and Oberhauser [2022] for a precise statement.

4.C.1 Marcus signature kernel

The definition of the signature kernel requires an initial algebraic setup. Let $\langle \cdot, \cdot \rangle_1$ be the Euclidean inner product on $\mathbb{R}^d$. Denote by $\otimes$ the standard outer product of vector spaces. For any $n \in \mathbb{N}$, we denote by $\langle \cdot, \cdot \rangle_n$ on $(\mathbb{R}^d)^{\otimes n}$ the canonical Hilbert-Schmidt inner product defined for any $\mathbf{a} = (a_1, \dots, a_n)$ and $\mathbf{b} = (b_1, \dots, b_n)$ in $(\mathbb{R}^d)^{\otimes n}$ as $\langle \mathbf{a}, \mathbf{b} \rangle_n = \prod_{i=1}^n \langle a_i, b_i \rangle_1$. The inner product $\langle \cdot, \cdot \rangle_n$ on $(\mathbb{R}^d)^{\otimes n}$ can then be extended by linearity to an inner product $\langle \cdot, \cdot \rangle$ on $\tilde{T}((\mathbb{R}^d))$ defined for any $\mathbf{a} = (1, a_1, \dots)$ and $\mathbf{b} = (1, b_1, \dots)$ in $\tilde{T}((\mathbb{R}^d))$ as $\langle \mathbf{a}, \mathbf{b} \rangle = 1 + \sum_{n=1}^\infty \langle a_n, b_n \rangle_n$.

To begin with, let $\mathcal{X} = D_1([0,T], \mathbb{R}^d)$. If $x \in \mathcal{X}$ is càdlàg path, we can define the *Marcus signature* in the spirit of Marcus SDEs [Marcus, 1978, 1981] as the signature of the *Marcus interpolation* of x. This interpolation, denoted by $\hat{x}$, is the continuous path on $[0,T]$ obtained from x by linearly traversing the jumps of x over added fictitious time $r > 0$ and then reparameterising so that the path runs over $[0,T]$ instead of $[0, T+r]$. The general construction is given in Appendix 4.A. If x is continuous, x and $\hat{x}$ coincide; thus, without any ambiguity, we can define the Marcus signature $S(x)$ of a general bounded variation càdlàg path as the tensor series described above, but replacing x with $\hat{x}$ (see also the definition in 4.A.2).

Since the signature is invariant to certain reparameterisations (Propisition 1), it is not an injective map. Injectivity is a crucial property required to ensure characteristicness of the resulting signature kernel that we will introduce next. One way of overcome this issue is to to augment a path x with a time coordinate resulting in the path $\tilde{x} = (x, t)$[4]. The Marcus signature kernel is then naturally defined as the map $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $k(x, y) = \langle S(\tilde{x}), S(\tilde{y}) \rangle$ for any $x, y \in \mathcal{X}$. As stated in Theorem 5, this kernel is universal on compact subsets $K \subset \mathcal{X}$ and, equivalently, characteristic to the space of regular Borel measures on K. However, these properties do not generalize to the whole space $C_b(\mathcal{X}, \mathbb{R})$ of bounded continuous functions from $\mathcal{X}$ to $\mathbb{R}$.

In Chevyrev and Oberhauser [2022] the authors address this issue in the case of continuous paths by introducing the so-called *robust signature*. They define a *tensor normalization* as a continuous injective map

$$\Lambda : \tilde{T}\left(\left(\mathbb{R}^d\right)\right) \to \left\{ \mathbf{a} \in \tilde{T}\left(\left(\mathbb{R}^d\right)\right) \mid \|\mathbf{a}\| \leq R \right\}$$

for some $R > 0$ and such that $\Lambda(\mathbf{a}) = (\mathbf{a}^0, \lambda(\mathbf{a}) a_1, \lambda(\mathbf{a})^2 a_2, \dots)$ for some $\lambda : \tilde{T}\left(\left(\mathbb{R}^d\right)\right) \to (0, \infty)$.

Now, let $p \in [1, 3)$ and take $C_0^1(\mathbb{R}^d)$ to be the space of absolutely continuous functions on $\mathbb{R}^d$. Recall that $\Omega_{0,p}^D(\mathbb{R}^d)$ is the closure of $C_0^1(\mathbb{R}^d)$ in $\Omega_p^D(\mathbb{R}^d)$ under the metric α_p.

[4] If $\mathbf{x}$ is a càdlàg rough path, this is done via a *Young pairing* which results in a càdlàg p-rough path, $\tilde{\mathbf{x}}$, where the first level is given by (x_t, t). For more information, we refer to Appendix 4.A.3.

Throughout we let $\mathcal{X} = \Omega^D_{0,p}(\mathbb{R}^d)$ be a metric space equipped with α_p. Naturally, we can then define the signature kernel on $\mathcal{X}$ by $k(\mathbf{x}, \mathbf{y}) = \langle S(\tilde{\mathbf{x}}), S(\tilde{\mathbf{y}})\rangle$ and, similarly, the robust signature kernel $k_\Lambda(\mathbf{x}, \mathbf{y}) = \langle \Lambda(S(\tilde{\mathbf{x}})), \Lambda(S(\tilde{\mathbf{y}}))\rangle$ where Λ is a tensor normalisation.

Theorem 5. *Let $p \geq 1$, Λ a tensor normalization, and $K \subset \mathcal{X}$ compact under α_p. Then,*

(i) The signature kernel k is universal to $\mathcal{F} = C(K, \mathbb{R})$ equipped with the uniform topology and characteristic to the dual $\mathcal{F}'$, the space of regular Borel measures on K.

(ii) The robust signature kernel k_Λ is universal to $\mathcal{F} = C_b(\mathcal{X}, \mathbb{R})$ equipped with the strict topology and charactersitic to the dual $\mathcal{F}'$, the space of all finite Borel measures on $\mathcal{X}$.

Proof of Theorem 5. Part (i) follows directly from the proof of Proposition 3.6 in Cuchiero et al. [2022]. For part (ii) we shall proof that the feature map $F = \Lambda \circ S$ is universal and characteristic. The result then follows from Proposition 29 in Chevyrev and Oberhauser [2022]. We start by defining $\mathcal{P} = \mathcal{X}/\sim_t$ where the equivalence relation $\sim_t$ is defined in Appendix 4.A.2. We equip $\mathcal{P}$ with the topology induced by the embedding $S : \mathcal{P} \to \tilde{T}((\mathbb{R}^{d+1}))$. By Proposition 1, F is a continuous and injective map from $\mathcal{P}$ into a bounded subset of $\tilde{T}((\mathbb{R}^{d+1}))$. Thus, $\mathcal{H} = \{\langle \ell, F\rangle \mid \ell \in T((\mathbb{R}^{d+1}))'\}$ is a subset of $\mathcal{F}$ that seperates points. Furthermore, since F takes values in the set of group-like elements, $\mathcal{H}$ is a subalgebra of $\mathcal{F}$ (under the shuffle product). It then follows from Theorem 7 and Theorem 9 in Chevyrev and Oberhauser [2022] that F is universal and charecteristic. The fact that $\mathcal{F}'$ is the space of all finite Borel measures on $\mathcal{X}$ is part (iii) of Theorem 9 in the same paper. Finally, as per Appendix 4.A.3, the map $\mathbf{x} \mapsto \tilde{\mathbf{x}}$ is a continuous and injective embedding of $\mathcal{X}$ into $\mathcal{P}$ from which the result then follows. □

With d_k denoting the MMD for a given kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, the following is a direct consequence of Theorem 5.

Corollary 1. *Let $p \geq 1$, Λ a tensor normalization, and $K \subset \mathcal{X}$ compact under α_p. Then, d_k is a metric on $\mathcal{M}(K)$ and d_{k_Λ} is a metric on $\mathcal{M}(\mathcal{X})$.*

4.D. Forward sensitivities for SLIF network

In the general SSNN model, Theorem 2 gives the following result.

Proposition 1. *Fix some weight $w_{ij} \in w$, a neuron $k \in [K]$ and let $\mathcal{G}^k_t$ denote the gradient of (v^k, i^k) wrt. w_{ij} at time t. Furthermore, define $\gamma : \{0, 1\} \to \mathbb{R}^2$ such that $\gamma_0 = (\mu_1, -\mu_2)w_{lk}$, $\gamma_1 = (\mu_1, 0)v_{reset}$, and let $\Gamma \in \mathbb{R}^{2\times 2}$ be the drift matrix in the inter-spike SDE of (v^k, i^k). Then,*

$$\mathcal{G}^k_t = e^{\Gamma(t-s)}\left(\mathcal{G}^k_s - \gamma_{\delta_{lk}}\partial_{w_{ij}} s + \delta_{il}\delta_{jk}e_2\right), \tag{4.D.1}$$

where $e_n \in \mathbb{R}_2$ is the n'th unit vector, l is the neuron in $\mathrm{Pa}_k \cup \{k\}$ with the most recent spike time before t, and we denote this spike time by s. If t is a spike time of neuron k it therefore follows that

$$\partial_{w_{ij}} t = \frac{\lambda(v^k_{t_{prev}})\partial_{w_{ij}} t_{prev} - \int_{t_{prev}}^t \nabla\lambda(v^k_r) e_1^T \mathcal{G}^k_r dr}{\lambda(v^k_t)}, \tag{4.D.2}$$

where t_{prev} is the previous spike time of neuron k. In the case of a deterministic SNN, formula (4.D.2) is replaced by

$$\partial_{w_{ij}} t = -\frac{e_1^T \mathcal{G}^k_t}{\mu_1(i^k_t - v^k_t)}. \tag{4.D.3}$$

Proof. Throughout we fix some $t > 0$ and let $s < t$ denote most recent event time preceding t with l the index of the neuron firing at time s. We define the process $dw_t = 0dt$ with $w_0 = w_{ij}$ and with a slight abuse of notation we shall write $y^k_t = (v^k_t, i^k_t, s^k_t, w_t)$. We will leave out the event index n for notational simplicity. Since y^k_t depends on y_s only through y^k_s and $\nabla\mathcal{T}_l$ is block diagonal, a direct consequence of eq. (4.3.7) is

$$\mathcal{G}^k_t = \begin{pmatrix} I & 0 \end{pmatrix} \partial_{y^k_s} y^k_t \left(\nabla\mathcal{T}^l_j(y^k_{s-})\partial_{w_{ij}} y^k_{s-} - \left(\mu(y^k_s) - \nabla\mathcal{T}^k_l(y^k_{s-})\mu(y^k_{s-})\right)\partial_{w_{ij}} s \right)$$

where $\mu(v,i,s,w) = (\mu_1(i-v), -\mu_2 i, \lambda(v), 0)$. If $l \in \mathrm{Pa}_k \cup \{k\}$, then $\mathcal{T}^k_l = \mathrm{id}$ and therefore $\mathcal{G}^k_t = \begin{pmatrix} I & 0 \end{pmatrix} \partial_{y^k_s} y^k_t \partial_{w_{ij}} y^k_s$. One can then reapply the formula above until $l \in \mathrm{Pa}_k \cup \{k\}$. By the flow property, it follows that we may assume without loss of generality that $l \in \mathrm{Pa}_k \cup \{k\}$. This leaves us with two cases. We define $z^k_t = (v^k_t, i^k_t)$ so that $\begin{pmatrix} I & 0 \end{pmatrix} \partial_{y^k_s} y^k_t = \partial_{z^k_s} z^k_t$ and $\partial_{w_{ij}} z^k_t = \mathcal{G}^k_t$. Furthermore, let $a = \delta_{il}\delta_{jk}$.

Case 1, $l \in \mathrm{Pa}_k$: In this case $\mathcal{T}^k_l(v,i,s,w) = (v, i + aw + (1-a)c, s, w)$ where c is a constant. As a result

$$\begin{aligned} \partial_{z^k_s} z^k_t \nabla\mathcal{T}^k_l(y^k_{s-})\partial y^k_{s-} &= \partial_{z^k_s} z^k_t \mathcal{G}^k_t + a\partial_{i^k_s} z^k_t, \\ \partial_{z^k_s} z^k_t \left(\mu(y^k_s) - \nabla\mathcal{T}^k_l(y^k_{s-})\mu(y^k_{s-})\right) &= \partial_{z^k_s} z^k_t \gamma_0, \end{aligned}$$

In total,

$$\mathcal{G}^k_t = \partial_{z^k_s} z^k_t \left(\mathcal{G}^k_t - \gamma_0 \partial_{w_{ij}} s + a e_2 \right).$$

Case 2, $l = k$: In this case $\mathcal{T}^k_l(v,i,s,w) = (v - v_{reset}, i, \log u - \alpha, w)$ so that

$$\begin{aligned} \partial_{z^k_s} z^k_t \nabla\mathcal{T}^k_l(y^k_{s-})\partial y^k_{s-} &= \partial_{z^k_s} z^k_t \mathcal{G}^k_t, \\ \partial_{z^k_s} z^k_t \left(\mu(y^k_s) - \nabla\mathcal{T}^k_l(y^k_{s-})\mu(y^k_{s-})\right) &= \partial_{z^k_s} z^k_t \gamma_1, \end{aligned}$$

and, thus,

$$\mathcal{G}^k_t = \partial_{z^k_s} z^k_t \mathcal{G}^k_t - \partial_{z^k_s} z^k_t \gamma_0 \partial_{w_{ij}} s.$$

Note that z^k_t is an Ornstein-Uhlenbeck process initialized at z^k_s and with drift and diffusion matrices

$$\Gamma = \begin{pmatrix} -\mu_1 & \mu_1 \\ 0 & -\mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix}.$$

As a result, we can directly compute $\partial_{z_s^k} z_t^k = e^{(t-s)\Gamma}$. This proves that eq. (4.D.1) holds. Eq. (4.D.2) then follows directly from (4.3.6) and the fact that $\mathcal{E}_k(y) = s^k$. □

From this the results of Section 4.4.4 follow since the terms $\partial_{w_{ij}} s$ vanish whenever s is the spike time of a neuron l that is not a descendant of neuron j. Thus, equation (4.D.1) only includes terms depending on the activity of the pre and post-synaptic neuron. In particular, there is no need to store the gradient path $\mathcal{G}_t^k$ for each combination of neuron k and synapse ij, but each neuron only needs to keep track of the paths for its incoming synapses. This reduces the memory requirements from the order of K^3 to only K^2 (which is needed anyway to store the weight matrix). In general, the gradient paths can be approximated by simply omitting the terms $\partial_{w_{ij}} s$.

4.E. Experiments

4.E.1 Input current estimation

For each combination of sample size and σ we sample a data set of spike trains using Algorithm 1 with $N = 3$, i.e., up until the first three spikes are generated. We use `diffrax` to solve the inter-Event SDE with a step size of 0.01 and the numerical solver is the simple Euler-Maruyama method. We then sample an initial guess $c \sim \text{Unif}([0.5, 2.5])$ and run stochastic gradient descent using the approach described in 4.4.1. That is, for each step, we generate a batch of the same size as the sample size and use d_k to compare the generated batch to the data. For each step we also compare the absolute error between the average spike time of the first three spikes of the generated sample to a hold a out test set of the same size as the sample. We use the RMSProp algorithm with a decay rate of 0.7 and a momentum of 0.3 which we found to work well in practice. The learning rate is 0.001. The experiment was run locally on CPU with an Apple M1 Pro chip with 8 cores and 32 GB of ram. The entire experiment took approximately 3-6 hours to run. For the exact details of this experiment we refer to the notebook `snnax/notebooks/single_neuron.ipynb` in the supplementary material.

4.E.2 Synaptic weight estimation

As above, for each sample size $D \in \{256, 512, 1024\}$ we sample a data set of spike trains using Algorithm 1 with $T = 1$ and with the same differential equation solver setup as above. Thus, in this case, the number of spikes varies across each sample path. The parameters are chosen as follows:

- $v_{reset} = 1.2$
- $\lambda(v) = \exp(5(v - 1)$
- $\mu = (6, 5)$
- $\sigma = I_2/4$

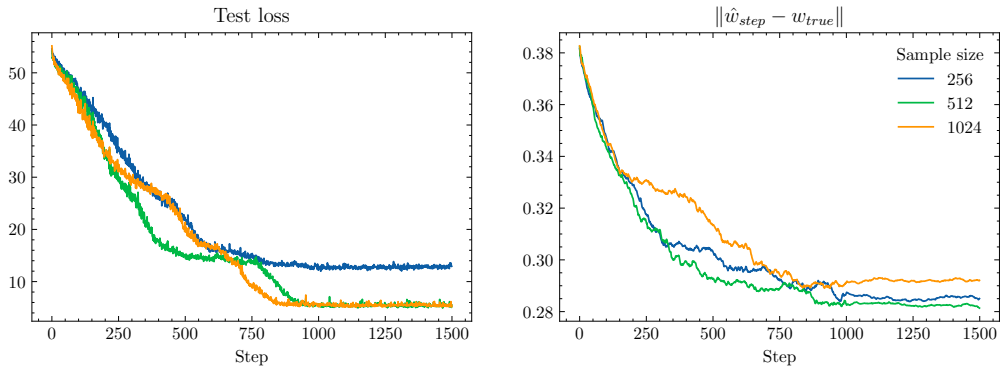

Figure 4.E.1: We estimate the synpatic weights w across three different sample sizes using the signature kernel MMD truncated at depth 3 and stochastic gradient descent with a batch size of 128. On the left we report the loss on a hold out test set. On the right is the mean absolute error between the entries of the currently estimated weight matrix $\hat{w}_{step}$ and the true weight matrix w_{true}.

For each sample size the data was generated using the same randomly sampled weight matrix w which represents a feed-forward network of the dimensions described in Section 4.4 and which was constructed as follows: for the weight matrix from layer l to layer $l+1$, say w^l, we sample each entry from Unif($[0.5, 1.5]$) and then normalize by $3/K_l$ where K_l is the number of neurons in layer l. The normalisation makes sure that the spike rate for the neurons in each layer is appropriate.

For each data set (each sample size) we then train a spiking neural net of the same network structure to match the observed spike trains. This is done using stochastic gradient descent with a batch size of $B = 128$ and by computing d_k on a generated batch and a batch sampled from the data set at each step. In order to avoid local minimums[5] we match the number of spikes between the generated spike trains and the ones sampled from the data set. Also, we sample from the data set without replacement so that we loop through the whole data set every D/B steps. We run RMSProp for 1500 steps with a momentum of 0.3 and a learning rate of 0.003 for the first 1000 steps and 0.001 for the last 500 steps.

This experiment was run in the cloud using Azure AI Machine Learning Studio on a NVIDIA Tesla V100 GPU with 6 cores and 112 GB of RAM. The entire experiment took around 12-16 hours to run. For the exact details we refer to the notebook `snnax/notebooks/spiking_neural_net.ipynb` in the supplementary material.

[5]Note that the loss landscape is inherently discontinuous since whenever the parameters are altered in such a way that an additional spike appears, the expected signature will jump.

Bibliography

R. P. Adams and D. J. MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

S. Aminikhanghahi and D. J. Cook. A survey of methods for time series change point detection. *Knowledge and information systems*, 51(2):339–367, 2017.

T. Anderson. Specification and misspecification in reduced rank regression. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 193–205, 2002a.

T. W. Anderson. Asymptotic distribution of the reduced rank regression estimator under general conditions. *The Annals of Statistics*, 27(4):1141–1154, 1999.

T. W. Anderson. Reduced rank regression in cointegrated models. *Journal of Econometrics*, 106(2):203–216, 2002b.

D. W. Andrews. Exactly median-unbiased estimation of first order autoregressive/unit root models. *Econometrica: Journal of the Econometric Society*, pages 139–165, 1993.

D. W. Andrews, X. Cheng, and P. Guggenberger. Generic results for establishing the asymptotic size of confidence sets and tests. *Journal of Econometrics*, 218(2):496–531, 2020.

F. Archontakis. An alternative proof of granger's representation theorem for i (1) systems through jordan matrices. *Journal of the Italian Statistical Society*, 7:111–127, 1998.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

L. Arnold. *Random Dynamical Systems*. Springer Monographs in Mathematics. Springer Berlin Heidelberg, 2013. ISBN 9783662128787.

M. Baktashmotlagh, M. Harandi, B. C. Lovell, and M. Salzmann. Discriminative non-linear stationary subspace analysis for video classification. *IEEE transactions on pattern analysis and machine intelligence*, 36(12):2353–2366, 2014.

N. S. Balke and T. B. Fomby. Threshold cointegration. *International economic review*, pages 627–645, 1997.

F. Bec and A. Rahbek. Vector equilibrium correction models with non-linear discontinuous adjustments. *The Econometrics Journal*, 7(2):628–651, 2004.

G. Bellec, F. Scherr, A. Subramoney, E. Hajek, D. Salaj, R. Legenstein, and W. Maass. A solution to the learning dilemma for recurrent networks of spiking neurons. *Nature communications*, 11(1):3625, 2020.

A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

D. H. Bernstein and B. Nielsen. Asymptotic theory for cointegration analysis when the cointegration rank is deficient. *Econometrics*, 7(1):6, 2019.

J. Y. Campbell and M. Yogo. Efficient tests of stock return predictability. *Journal of financial economics*, 81(1):27–60, 2006.

T. Cass and C. Salvi. Lecture notes on rough paths and applications to machine learning. *arXiv preprint arXiv:2404.06583*, 2024.

G. Cavaliere, A. Rahbek, and A. R. Taylor. Bootstrap determination of the co-integration rank in vector autoregressive models. *Econometrica*, 80(4):1721–1740, 2012.

G. Cavaliere, A. Rahbek, and A. Robert Taylor. Bootstrap determination of the co-integration rank in var models with unrestricted deterministic components. *Journal of Time Series Analysis*, 36(3):272–289, 2015.

M. Cenedese, J. Axås, B. Bäuerlein, K. Avila, and G. Haller. Data-driven modeling and prediction of non-linearizable dynamics via spectral submanifolds. *Nature communications*, 13(1):872, 2022.

R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

R. T. Chen, B. Amos, and M. Nickel. Learning neural event functions for ordinary differential equations. In *International Conference on Learning Representations*, 2020.

B.-E. Chérief-Abdellatif and P. Alquier. Finite sample properties of parametric mmd estimation: robustness to misspecification and dependence. *Bernoulli*, 28(1):181–213, 2022.

I. Chevyrev and P. K. Friz. Canonical rdes and general semimartingales as rough paths. *The Annals of Probability*, 47(1):420–463, 2019.

I. Chevyrev and T. Lyons. Characteristic functions of measures on geometric rough paths. *The Annals of Probability*, 44(6):4049 – 4082, 2016.

I. Chevyrev and H. Oberhauser. Signature moments to characterize laws of stochastic processes. *Journal of Machine Learning Research*, 23(176):1–42, 2022.

S. Choi, A. Cichocki, H.-M. Park, and S.-Y. Lee. Blind source separation and independent component analysis: A review. *Neural Information Processing-Letters and Reviews*, 6(1):1–57, 2005.

N. M. Cirone, M. Lemercier, and C. Salvi. Neural signature kernels as infinite-width-depth-limits of controlled resnets. In *International Conference on Machine Learning*, pages 25358–25425. PMLR, 2023.

T. Cochrane, P. Foster, V. Chhabra, M. Lemercier, T. Lyons, and C. Salvi. Sk-tree: a systematic malware detection algorithm on streaming trees via the signature kernel. In *2021 IEEE international conference on cyber security and resilience (CSR)*, pages 35–40. IEEE, 2021.

S. Corner, C. Sandu, and A. Sandu. Modeling and sensitivity analysis methodology for hybrid dynamical system. *Nonlinear Analysis: Hybrid Systems*, 31:19–40, 2019.

S. Corner, A. Sandu, and C. Sandu. Adjoint sensitivity analysis of hybrid multibody dynamical systems. *Multibody System Dynamics*, 49:395–420, 2020.

C. Cuchiero, F. Primavera, and S. Svaluto-Ferro. Universal approximation theorems for continuous functions of c\adl\ag paths and l\'evy-type signature models. *arXiv preprint arXiv:2208.02293*, 2022.

C. Cuny, J. Dedecker, and F. Merlevède. Rates of convergence in invariance principles for random walks on linear groups via martingale methods. *Transactions of the American Mathematical Society*, 374(1):137–174, 2021.

R. M. De Jong. Weak laws of large numbers for dependent random variables. *Annales d'Économie et de Statistique*, pages 209–225, 1998.

H. Dehling and M. Wendler. Central limit theorem and the bootstrap for u-statistics of strongly mixing data. *Journal of Multivariate Analysis*, 101(1):126–137, 2010.

J. J. Dolado and H. Lütkepohl. Making wald tests work for cointegrated var systems. *Econometric reviews*, 15(4):369–386, 1996.

P. Doukhan. *Mixing: properties and examples*, volume 85. Springer Science & Business Media, 2012.

J. Duffy and J. Simons. Cointegration without unit roots. Technical report, Faculty of Economics, University of Cambridge, 2023.

J. A. Duffy, S. Mavroeidis, and S. Wycherley. Cointegration with occasionally binding constraints. *arXiv preprint arXiv:2211.09604*, 2022.

G. Elliott. On the robustness of cointegration methods when regressors almost have unit roots. *Econometrica*, 66(1):149–158, 1998.

G. Elliott and J. H. Stock. Inference in time series regression when the order of integration of a regressor is unknown. *Econometric theory*, 10(3-4):672–700, 1994.

G. Elliott, U. K. Muller, and M. W. Watson. Nearly optimal tests when a nuisance parameter is present under the null hypothesis. *Econometrica*, 83(2):771–811, 2015.

R. Engle and C. Granger. Cointegration and error correction - representation, estimation, and testing. *Econometrica*, 55(2):251–276, 3 1987. ISSN 0012-9682.

A. Escribano. Nonlinear error correction: The case of money demand in the united kingdom (1878–2000). *Macroeconomic Dynamics*, 8(1):76–116, 2004.

A. Fermanian, T. Lyons, J. Morrill, and C. Salvi. New directions in the applications of rough path theory. *IEEE BITS the Information Theory Magazine*, 2023.

M. Franchi and S. Johansen. Improved inference on cointegrating vectors in the presence of a near unit root using adjusted quantiles. *Econometrics*, 5(2):25, 2017.

P. Friz and A. Shekhar. General rough integration, lévy rough paths and a lévy–kintchine-type formula. *Annals of probability: An official journal of the Institute of Mathematical Statistics*, 45(4):2707–2765, 2017.

P. K. Friz and M. Hairer. *A course on rough paths*. Springer, 2020.

P. K. Friz and N. B. Victoir. *Multidimensional stochastic processes as rough paths: theory and applications*, volume 120. Cambridge University Press, 2010.

P. K. Friz and H. Zhang. Differential equations driven by rough paths with jumps. *Journal of Differential Equations*, 264(10):6226–6301, 2018.

D. Garreau and S. Arlot. Consistent change-point detection with kernels. *Electronic Journal of Statistics*, 12(2):4440 – 4486, 2018.

W. Gerstner and W. M. Kistler. *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge university press, 2002.

C. W. Granger and P. Newbold. Spurious regressions in econometrics. *Journal of econometrics*, 2(2):111–120, 1974.

J. Gygax and F. Zenke. Elucidating the theoretical underpinnings of surrogate gradient learning in spiking neural networks. *arXiv preprint arXiv:2404.14964*, 2024.

B. Hambly and T. Lyons. Uniqueness for the signature of a path of bounded variation and the reduced path group. *Annals of Mathematics*, pages 109–167, 2010.

B. E. Hansen. The grid bootstrap and the autoregressive model. *Review of Economics and Statistics*, 81(4):594–607, 1999.

B. E. Hansen. Averaging estimators for autoregressions with a near unit root. *Journal of Econometrics*, 158(1):142–155, 2010.

Z. Harchaoui, E. Moulines, and F. Bach. Kernel change-point analysis. *Advances in neural information processing systems*, 21, 2008.

T. A. Henzinger. The theory of hybrid automata. In *Proceedings 11th Annual IEEE Symposium on Logic in Computer Science*, pages 278–292. IEEE, 1996.

M. Höglund, E. Ferrucci, C. Hernández, A. M. Gonzalez, C. Salvi, L. Sánchez-Betancourt, and Y. Zhang. A neural rde approach for continuous-time non-markovian stochastic control problems. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, 2023.

C. Holberg. Stationary embeddings: A nonlinear generalization of cointegration, 2024. Working paper.

C. Holberg and S. Ditlevsen. Weighted reduced rank estimators under cointegration rank uncertainty. *Scandinavian Journal of Statistics*, 2024a. To appear.

C. Holberg and S. Ditlevsen. Uniform inference for cointegrated vector autoregressive processes. *Journal of Econometrics*, 2024b. To appear.

C. Holberg and C. Salvi. Exact gradients for stochastic spiking neural networks driven by rough signals. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

A. Hyvarinen and H. Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

A. Hyvarinen and H. Morioka. Nonlinear ICA of Temporally Dependent Stationary Sources. In A. Singh and J. Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 460–469. PMLR, 20–22 Apr 2017.

A. Hyvarinen, H. Sasaki, and R. Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019.

Z. Issa and B. Horvath. Non-parametric online market regime detection and regime clustering for multidimensional and path-dependent data structures, 2023.

Z. Issa, B. Horvath, M. Lemercier, and C. Salvi. Non-adversarial training of neural sdes with signature kernel scores. *Advances in Neural Information Processing Systems*, 2023a.

Z. Issa, B. Horvath, M. Lemercier, and C. Salvi. Non-adversarial training of neural sdes with signature kernel scores. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 11102–11126. Curran Associates, Inc., 2023b.

A. J. Izenman. Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, 5(2):248–264, 1975.

H. Jang and O. Simeone. Multisample online learning for probabilistic spiking neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(5):2034–2044, 2022.

M. Jansson and M. J. Moreira. Optimal inference in regression models with nearly integrated regressors. *Econometrica*, 74(3):681–714, 2006.

J. Jia and A. R. Benson. Neural jump stochastic differential equations. *Advances in Neural Information Processing Systems*, 32, 2019.

D. Jimenez Rezende and W. Gerstner. Stochastic variational learning in recurrent spiking networks. *Frontiers in computational neuroscience*, 8:38, 2014.

S. Johansen. Statistical analysis of cointegration vectors. *Journal of economic dynamics and control*, 12(2-3):231–254, 1988.

S. Johansen. Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. *Econometrica: journal of the Econometric Society*, pages 1551–1580, 1991.

S. Johansen. *Likelihood-based inference in cointegrated vector autoregressive models*. Oxford University Press on Demand, 1995.

H. Kaido, F. Molinari, J. Stoye, and M. Thirkettle. Calibrated projection in matlab: Users' manual. *arXiv preprint arXiv:1710.09707*, 2017.

H. Kaido, F. Molinari, and J. Stoye. Confidence intervals for projections of partially identified parameters. *Econometrica*, 87(4):1397–1432, 2019.

H. Kajino. A differentiable point process with its application to spiking neural networks. In *International Conference on Machine Learning*, pages 5226–5235. PMLR, 2021.

I. Karatzas and S. Shreve. *Brownian motion and stochastic calculus*, volume 113. Springer Science & Business Media, 1991.

H. A. Karlsen and D. Tjøstheim. Nonparametric estimation in null recurrent time series. *Annals of Statistics*, pages 372–416, 2001.

H. A. Karlsen, T. Myklebust, and D. Tjøstheim. Nonparametric estimation in a nonlinear cointegration type model. *The Annals of Statistics*, pages 252–299, 2007.

M. Kasy. Uniformity and the delta method. *Journal of Econometric Methods*, 8(1), 2019.

T. Kato. *Perturbation theory for linear operators*, volume 132. Springer Science & Business Media, 2013.

M. Kessler and A. Rahbek. Asymptotic likelihood based inference for co-integrated homogenous gaussian diffusions. *Scandinavian Journal of Statistics*, 28(3):455–470, 2001.

M. Kessler and A. Rahbek. Identification and inference for multivariate cointegrated and ergodic gaussian diffusions. *Statistical inference for stochastic processes*, 7:137–151, 2004.

I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International conference on artificial intelligence and statistics*, pages 2207–2217. PMLR, 2020a.

I. Khemakhem, R. Monti, D. Kingma, and A. Hyvarinen. Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ica. *Advances in Neural Information Processing Systems*, 33:12768–12778, 2020b.

P. Kidger. *On Neural Differential Equations*. PhD thesis, University of Oxford, 2021.

P. Kidger, J. Foster, X. Li, and T. J. Lyons. Neural sdes as infinite-dimensional gans. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5453–5463. PMLR, 18–24 Jul 2021.

F. J. Király and H. Oberhauser. Kernels for sequentially ordered data. *Journal of Machine Learning Research*, 20(31):1–45, 2019.

G. Koop, R. Strachan, H. Van Dijk, and M. Villani. Bayesian approaches to cointegration. *Research Papers in Economics*, pages 871–898, 2006.

A. Kostakis, T. Magdalinos, and M. P. Stamatogiannis. Robust econometric inference for stock return predictability. *The Review of Financial Studies*, 28(5):1506–1553, 2015.

S. Krantz and H. Parks. *A Primer of Real Analytic Functions*. A Primer of Real Analytic Functions. Birkhäuser Boston, 2002. ISBN 9780817642648.

D. Kristensen and A. Rahbek. Likelihood-based inference in nonlinear error-correction models. *Journal of Econometrics*, 158, 12 2007.

J. Krystul and H. Blom. Generalised stochastic hybrid processes as strong solutions of stochastic differential equations. *Hybridge report D*, 2, 2005.

J. Krystul, H. A. Blom, and A. Bagchi. Stochastic differential equations on hybrid state spaces. *Stochastic Hybrid Systems*, 24(15-45):170, 2006.

P. Lansky and S. Ditlevsen. A review of the methods for signal estimation in stochastic diffusion leaky integrate-and-fire neuronal models. *Biological cybernetics*, 99(4-5):253–262, 2008.

D. Lee and H. Oberhauser. The signature kernel. *arXiv preprint arXiv:2305.04625*, 2023.

J. H. Lee, S. Haghighatshoar, and A. Karbasi. Exact gradient computation for spiking neural networks via forward propagation. In *International Conference on Artificial Intelligence and Statistics*, pages 1812–1831. PMLR, 2023.

Y.-S. Lee, T.-H. Kim, and P. Newbold. Spurious nonlinear regressions in econometrics. *Economics Letters*, 87(3):301–306, 2005.

M. Lemercier, C. Salvi, T. Damoulas, E. Bonilla, and T. Lyons. Distribution regression for sequential data. In *International Conference on Artificial Intelligence and Statistics*, pages 3754–3762. PMLR, 2021.

M. Levakova and S. Ditlevsen. Penalization methods in fitting high-dimensional cointegrated VAR models: a review. *International Statistical Review*, page 220621, 2023. To appear.

M. Levakova, J. H. Christensen, and S. Ditlevsen. Classification of brain states that predicts future performance in visual tasks based on co-integration analysis of eeg data. *Royal Society Open Science*, 9(11):220621, 2022.

D. Li, D. Tjostheim, and J. Gao. Estimation in nonlinear regression with harris recurrent markov chains. *Annals of Statistics*, 44(5):1957–1987, 2016.

S. Li, Y. Xie, H. Dai, and L. Song. M-statistic for kernel change-point detection. *Advances in Neural Information Processing Systems*, 28, 2015.

X. Li, T.-K. L. Wong, R. T. Chen, and D. K. Duvenaud. Scalable gradients and variational inference for stochastic differential equations. In *Symposium on Advances in Approximate Bayesian Inference*, pages 1–28. PMLR, 2020.

L. Lieb and S. Smeekes. Inference for impulse responses under model uncertainty. WorkingPaper 022, Maastricht University, Graduate School of Business and Economics, Netherlands, Oct. 2017.

A. R. Lundborg, R. D. Shah, and J. Peters. Conditional independence testing in Hilbert spaces with applications to functional data analysis. *Journal of the Royal Statistical Society, Series B*, 84(5):1821–1850, 2022.

H. Lütkepohl. *New introduction to multiple time series analysis.* Springer Science & Business Media, 2005.

J. Lygeros and M. Prandini. Stochastic hybrid systems: a powerful framework for complex, large scale applications. *European Journal of Control*, 16(6):583–594, 2010.

T. Lyons. Rough paths, signatures and the modelling of functions on streams. *arXiv preprint arXiv:1405.4537*, 2014.

T. J. Lyons. Differential equations driven by rough signals. *Revista Matemática Iberoamericana*, 14(2):215–310, 1998.

T. J. Lyons, M. Caruana, and T. Lévy. *Differential equations driven by rough paths*. Springer, 2007.

G. Ma, R. Yan, and H. Tang. Exploiting noise as a resource for computation and learning in spiking neural networks. *Patterns*, 4(10), 2023.

T. Magdalinos and P. C. B. Phillips. Econometric inference in matrix vicinities of unity and stationarity. *Working Paper*, 2020.

J. R. Magnus and H. Neudecker. The commutation matrix: some properties and applications. *The Annals of Statistics*, 7(2):381–394, 1979.

J. R. Magnus and H. Neudecker. *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons, 2019.

G. Manten, C. Casolo, E. Ferrucci, S. W. Mogensen, C. Salvi, and N. Kilbertus. Signature kernel conditional independence tests in causal discovery for stochastic processes. *arXiv preprint arXiv:2402.18477*, 2024.

S. Marcus. Modeling and analysis of stochastic differential equations driven by point processes. *IEEE Transactions on Information theory*, 24(2):164–172, 1978.

S. I. Marcus. Modeling and approximation of stochastic differential equations driven by semimartingales. *Stochastics: An International Journal of Probability and Stochastic Processes*, 4(3):223–245, 1981.

A. Mikusheva. Uniform inference in autoregressive models. *Econometrica*, 75(5):1411–1452, 2007.

A. Mikusheva. One-dimensional inference in autoregressive models with the potential presence of a unit root. *Econometrica*, 80(1):173–212, 2012.

J. Morrill, C. Salvi, P. Kidger, and J. Foster. Neural rough differential equations for long time series. In *International Conference on Machine Learning*, pages 7829–7838. PMLR, 2021.

K. Muandet, K. Fukumizu, B. Sriperumbudur, B. Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

E. O. Neftci, H. Mostafa, and F. Zenke. Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 36(6):51–63, 2019.

H. Neudecker. The kronecker matrix product and some of its applications in econometrics. *Statistica Neerlandica*, 22(1):69–82, 1968.

B. Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

A. Onatski and C. Wang. Alternative asymptotics for cointegration tests in large vars. *Econometrica*, 86(4):1465–1478, 2018.

A. Pakniyat and P. E. Caines. On the stochastic minimum principle for hybrid systems. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 1139–1144. IEEE, 2016.

A. Pannier and C. Salvi. A path-dependent pde solver based on signature kernels. *arXiv preprint arXiv:2403.11738*, 2024.

J. Y. Park and P. C. Phillips. Asymptotics for nonlinear transformations of integrated time series. *Econometric Theory*, 15(3):269–298, 1999.

J. Y. Park and P. C. Phillips. Nonlinear regressions with integrated time series. *Econometrica*, 69(1):117–161, 2001.

J.-P. Pfister, T. Toyoizumi, D. Barber, and W. Gerstner. Optimal spike-timing-dependent plasticity for precise action potential firing in supervised learning. *Neural computation*, 18(6):1318–1348, 2006.

P. C. Phillips. Understanding spurious regressions in econometrics. *Journal of econometrics*, 33(3):311–340, 1986.

P. C. Phillips. Regression theory for near-integrated time series. *Econometrica: Journal of the Econometric Society*, pages 1021–1043, 1988.

P. C. Phillips. Optimal inference in cointegrated systems. *Econometrica: Journal of the Econometric Society*, pages 283–306, 1991.

P. C. Phillips. Local limit theory and spurious nonparametric regression. *Econometric Theory*, 25(6):1466–1497, 2009.

P. C. Phillips. On confidence intervals for autoregressive roots and predictive regression. *Econometrica*, 82(3):1177–1195, 2014.

P. C. Phillips and J. H. Lee. Predictive regression under various degrees of persistence and robust long-horizon regression. *Journal of Econometrics*, 177(2):250–264, 2013.

P. C. Phillips and J. H. Lee. Limit theory for vars with mixed roots near unity. *Econometric Reviews*, 34(6-10):1035–1056, 2015.

P. C. Phillips and T. Magdalinos. Limit theory for moderate deviations from a unit root. *Journal of Econometrics*, 136(1):115–130, 2007.

P. C. Phillips, T. Magdalinos, et al. Econometric inference in the vicinity of unity. *Singapore Management University, CoFie Working Paper*, 7, 2009.

P. Pilipovic, A. Samson, and S. Ditlevsen. Parameter estimation in nonlinear multivariate stochastic differential equations based on splitting schemes. *The Annals of Statistics*, 52(2):842 – 867, 2024a.

P. Pilipovic, A. Samson, and S. Ditlevsen. Strang splitting for parametric inference in second-order stochastic differential equations. *arXiv preprint arXiv:2405.03606*, 2024b.

D. Pollard. *Convergence of Stochastic Processes*. Springer Series in Statistics. Springer New York, 1984.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

L. C. G. Rogers and D. Williams. *Diffusions, Markov Processes, and Martingales*. Cambridge Mathematical Library. Cambridge University Press, 2 edition, 2000.

R. Roy. Asymptotic covariance structure of serial correlations in multivariate time series. *Biometrika*, 76(4):824–827, 1989.

C. Salvi. *Rough paths, kernels, differential equations and an algebra of functions on streams*. PhD thesis, University of Oxford, 2021.

C. Salvi, T. Cass, J. Foster, T. Lyons, and W. Y. The signature kernel is the solution of a Goursat PDE. *SIAM Journal on Mathematics of Data Science*, 3(3):873–899, 2021a.

C. Salvi, M. Lemercier, T. Cass, E. V. Bonilla, T. Damoulas, and T. J. Lyons. Siggpde: Scaling sparse gaussian processes on sequential data. In *International Conference on Machine Learning*, pages 6233–6242. PMLR, 2021b.

C. Salvi, M. Lemercier, C. Liu, B. Horvath, T. Damoulas, and T. Lyons. Higher order kernel mean embeddings to capture filtrations of stochastic processes. *Advances in Neural Information Processing Systems*, 34:16635–16647, 2021c.

A. Schell and H. Oberhauser. Nonlinear independent component analysis for discrete-time and continuous-time signals. *The Annals of Statistics*, 51(2):487–518, 2023.

B. Schölkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

J. Stærk-Østergaard, A. Rahbek, and S. Ditlevsen. High-dimensional cointegration and Kuramoto inspired systems. *SIAM Journal on Applied Dynamical Systems*, pages 1–20, 2023. To appear.

R. R. Sundararajan and M. Pourahmadi. Stationary subspace analysis of nonstationary processes. *Journal of Time Series Analysis*, 39(3):338–355, 2018.

D. Tjøstheim. Some notes on nonlinear cointegration: A partial review with some novel perspectives. *Econometric Reviews*, 39(7):655–673, 2020.

H. Y. Toda and T. Yamamoto. Statistical inference in vector autoregressions with possibly integrated processes. *Journal of econometrics*, 66(1-2):225–250, 1995.

C. Toth, H. Oberhauser, and Z. Szabo. Random fourier signature features. *arXiv preprint arXiv:2311.12214*, 2023.

C. Truong, L. Oudre, and N. Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020.

Y. Tu and Y. Wang. Spurious functional-coefficient regression models and robust inference with marginal integration. *Journal of Econometrics*, 229(2):396–421, 2022.

A. W. Van der Vaart. Time series. *VU University Amsterdam, lecture notes*, 2010.

P. Von Bünau, F. C. Meinecke, F. C. Király, and K.-R. Müller. Finding stationary subspaces in multivariate time series. *Physical review letters*, 103(21):214101, 2009.

J. Von Kügelgen, Y. Sharma, L. Gresele, W. Brendel, B. Schölkopf, M. Besserve, and F. Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021.

J. A. Wellner. *Empirical Processes: Theory and Applications*. Delft Technical University, 2005.

T. C. Wunderlich and C. Pehle. Event-based backpropagation can compute exact gradients for spiking neural networks. *Scientific Reports*, 11(1):12829, 2021.

M. Xiao, Q. Meng, Z. Zhang, D. He, and Z. Lin. Online training through time for spiking neural networks. *Advances in neural information processing systems*, 35:20717–20730, 2022.

Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

F. Zenke and T. P. Vogels. The remarkable robustness of surrogate gradient learning for instilling complex function in spiking neural networks. *Neural computation*, 33(4): 899–925, 2021.

R. Zhang, P. Robinson, and Q. Yao. Identifying cointegration by eigenanalysis. *Journal of the American Statistical Association*, 114(526):916–927, 2019.