Statistical Methods for Neural Data

Cointegration Analysis of Coupled Neurons & Generalized Linear Models for Spike Train Data

PhD thesis by
Jacob Østergaard

Department of Mathematical Sciences
University of Copenhagen
Denmark

The PhD School of Science - The Faculty of Science - University of Copenhagen

Jacob Østergaard

Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
DK-2100 København Ø
Denmark

ostergaard@math.ku.dk

This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen, December 2017.

ACADEMIC ADVISORS:
Susanne Ditlevsen
University of Copenhagen, Denmark

Anders Rahbek
University of Copenhagen, Denmark

ASSESSMENT COMMITTEE:
Niels Richard Hansen (chair)
University of Copenhagen, Denmark

Rainer Dahlhaus
Universität Heidelberg, Germany

Vincent Rivoirard
Université Paris Dauphine, France

ISBN 978-87-7078-903-5

ABSTRACT

Some of the most captivating questions in the history of science concerns the functions of the human brain and the subject has attracted researchers and philosophers for centuries. Recent advances in laboratory technology has enabled us to look further into the internal microscopic components of the brain than ever before. Neuroscience, as a purely scientific discipline, is relatively new compared to it's basic components of mathematics, physics, chemistry, and physiology. However, the current rate of experimental discoveries in neuroscience calls for new advances in analytical tools to better understand the biological processes that occur in the brain.

This thesis aims to explore new statistical models for neural data and their usefulness in analyzing experimental data. The thesis consist of two parts, one that concerns neural networks and how these can be interpreted as a *cointegrated* system and one that examines how the class of *Generalized Linear Models* can be used to decode specific behaviors of simulated neurons. Part one introduces the concept of cointegration and demonstrates how a network can be analyzed by interpreting the system as a cointegrated process. This work is then extended from a small 3-dimensional system to a high-dimensional setting and includes a discussion of future possibilities for network analysis using these techniques. Part two opens with a demonstration of how Generalized Linear Models can be designed for spike train data and how varying patterns of different neurons are captured by this class of statistical models. Part two then continues with a specialized model aimed at capturing a specific type of behavior known as "bursting".

In the age of big data and artificial intelligence, two major themes related to neuroscience present themselves. The first is how to cope with the rapidly increasing data collection from laboratory experiments and (very) high-dimensional interacting systems. This occurs partially due to an increased interest in neuroscience, as well as the introduction of new measuring equipment. The second is the motivation for a continuously deeper understandning of the human brain. There are still countless unanswered questions regarding this biological mechanism. In order to further understand causes of neural diseases as well as continued development of artificial intelligence, these questions are important to study. Ultimately, they should lead us to a better intuition regarding the question: "how does intelligence work"?

RESUME PÅ DANSK

Nogle af de mest fascinerende spørgsmål gennem videnskabens historie handler om hvordan den menneskelige hjerne fungerer, og emnet har optaget forskere og filosoffer i århundreder. Den teknologiske udvikling af laboratorie udstyr har gjort det muligt at vi idag kan se helt ind i hjernens inderste mikroskopiske dele. Neuroscience er et relativt nyt område indenfor videnskaben, når man sammenligner med de klassiske videnskaber det bygger på: matematik, fysik, kemi og fysiologi. Ikke desto mindre gør det nuværende tempo af eksperimentelle opdagelser indenfor feltet, at der er et øget behov for nye værktøjer til at hjælpe os med at forstå de biologiske processer der foregår i hjernen.

Denne afhandling undersøger nye statistiske modeller for neurale data og deres anvendelse til at analysere eksperimentelle data. Afhandlingen består af to dele hvoraf den første del handler om hvordan neurale netværk kan fortolkes som et *kointegreret* system, og den anden del undersøger hvordan modelklassen *Generaliserede Lineære Modeller* kan udnyttes til at afkode specifikke mønstre i simulerede neuroner. Første del introducerer konceptet kointegration og demonstrerer herefter hvordan et netværk kan analyseres som en kointegreret process. Derefter udvides ideen fra at se på et mindre 3-dimensionelt system til et høj-dimensionalt, sammen med en diskussion af nye muligheder for at analysere netværk med disse værktøjer. Anden del starter med en demonstration af hvordan Generaliserede Lineære Modeller kan bruges til at analysere "spike train" data og hvordan disse modeller kan beskrive varierende mønstre fra forskellige typer af neuroner. Anden del fortsætter derefter med en specialiseret model der har til formål at beskrive et specifik mønster, kendt som "bursting".

I en tidsalder med stor interesse for big data og kunstig intelligens, er der særligt to spørgsmål der relaterer sig til neuroscience. Det første drejer sig om at håndtere de hurtigt voksende mængder af data af (meget) høj-dimensional karakter. Dette er en konsekvens af både nyt laboratorie udstyr, samt en voksende interesse i neuroscience fra forskere. Det andet berører motivationen for en gradvis dybere forståelse af den menneskelige hjerne. Der er stadig utallige ubesvarede spørgsmål omkring hjernens biologiske funktioner, og for bedre at kunne forstå årsagen til fejl i nervesystemet samt videreudvikling af kunstig intelligens er disse spørgsmål vigtige at studere. I sidste ende vil det forhåbentligt lede os til en bedre forståelse for spørgsmålet: "hvordan fungerer intelligens"?

ACKNOWLEDGEMENTS

First and foremost I would like to thank my two supervisors Susanne Ditlevsen and Anders Rahbek with whom I have shared many lively discussions on the matters of this thesis. Both have been of invaluable assistance and have on numerous occasions helped to clarify some of my convoluted ideas and intuition. I would also like to thank my hosts and collaborators in Boston, Uri Eden and Mark Kramer. During my time there I had many enlightening discussions with both of them and their research group at Boston University. I would also like to thank Søren Johansen for pointing out interesting facts related to my research and intuition, regarding the link between cointegration and neuroscience. I would also like to thank the PhD secretary the math department in Copenhagen, Nina Weisse, for kind support on formal matters during the project. A thank you goes out to colleagues at the math department for interesting discussions related not only to mathematics and statistics, but also worldly matters, as well as helping to create a warm and engaging work environment. A special thanks goes out to Rainer Dahlhaus for his compelling work that led to the realization of this project. Finally I would like to thank my family and friends for support, especially Troels Mikkelsen who initially pointed me in the direction of Susanne Ditlevsen when I longed for a return to academia from the industry. And of course my warmest thank you to my fiancé, Laila Stærk, for supporting my decision to write a PhD thesis and the process itself and our daughter Freja, who was very considerate when coming into this world at the very final stage of this project.

"There are two kinds of statistics, the kind you look up and the kind you make up."

– Rex Stout

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Background

My exploration of statistical methods for neural data came about rather coincidentally when I stumbled upon Susanne Ditlevsen a warm July afternoon in 2014. At the time I was employed as a quantitative researcher in a financial institution, but felt that it was time for a change. Looking back, I had no perspective on neuroscience, but I had set up a meeting with Susanne regarding a possible return to academia as a PhD student at the University of Copenhagen. I had no clear expectations regarding the outcome of the meeting, but recall that I managed to mention both *dynamical systems* and *cointegration* in a sentence outlining my research interests. From then on the gears were in motion and 6 months later I started a project with the aim of researching cointegration analysis of dynamical systems.

I have now spent the past three years investigating the idea of, analyzing multivariate interacting dynamical processes, of a physiologically meaningful nature, using cointegration theory. This journey has been partially successful in establishing some connections, but it also led me to a series of new interesting questions regarding the application of cointegration analysis within neuroscience. Transitioning from the financial industry with a head dive into statistical neuroscience was a challenging move, but it rewarded me with fruitful experiences, including a seven month excursion to Boston where I had a chance to work with professors Uri Eden and Mark Kramer. This research stay gave me an opportunity to expand my horizon of statistical methods for neural data and work in a different environment than I was used to. Here I worked on analysis of spike train data with generalized linear models, a subject almost orthogonal to my work in Copenhagen. However, it helped me form some inspiring ideas on future directions of cointegration research as a valuable toolset for neuroscience.

I hope that my efforts over these years, which have only scratched the surface, will aid in an increased interest in cointegration applied outside of econometrics and that researchers will discover it as valuable theory to help explain more of one of the most fascinating biological organisms we know, the brain.

1.2 Foundations

The age of exploration was one of the great periods of enlightenment in human history. As the globe became charted and remote areas were discovered, humankind gradually increased their understanding of the world they were living in. In a similar manner today, we are discovering more and more areas of the brain, an organism which has fascinated researchers for centuries. New laboratory technologies and methods enable researchers to measure brain activity in animals and humans on a much more detailed level than ever before, much like advancements in ship construction enabled circumnavigations in the 16th century. However, also crucial to these explorations were theoretical discoveries in astronomy and navigation, and to fully explore the potentials of our brain, advancing the theoretical understanding of the fundamental building blocks, the neurons, is of undeniable importance. Fortunately, neuroscience have already been a subject of interest for more than a century and modern ideas and conclusions build upon decades of considerations and practical experiments on how the brain actually work.

In the following we briefly review some of the most fundmental concepts in mathematical neuroscience to set the stage for the theoretical introductions in this thesis.

1.2.1 What is a Neuron

The brain is in essence, a coupled network of billions of neurons[1]. Each of these consist of a cell body, the soma, which contains dendrites that pick up signals, electrical impulses, from other neurons. These signals are propagated through the axon to the "tail" of the neuron which ends in synapses that connect to other nearby neurons dendrites, creating a connected network of neurons. A single neuron cell is drawn in Figure 1.1.

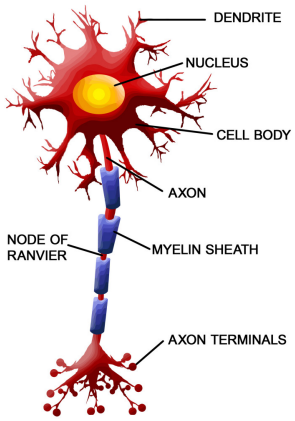

FIGURE 1.1: A neuron, image from *interactive-biology.com*. The dendrites receive input from nearby cells, propagate through the axon to the synapses, here termed *axon terminals*, which distribute the signal further in the information chain, to nearby cells.

[1]Some uncertainty persist in the community on exactly how many billions of neurons. A commonly cited number is 100 billion, such as in Izhikevich (2010), but Herculano-Houzel (2009) argue that it is more accurately around 86 billion. Still quite a few neurons!

1.2.2 The Action Potential

Neurons receive and transmit signals as electrical impulses which are controlled by *ionic currents* on both sides of the cell membrane. When the neuron is resting it has a charge of around -65 mV (Izhikevich, 2010) termed the *membrane potential*, but when the *ion channels* open, the currents flow in and out of the cell, thereby *depolarizing* it by increasing the charge inside the cell. When the depolarization hits a neuron specific threshold, the neuron *fires* by propagating a signal called the *action potential*, through the axon to the synapses. Immediately following this, the neuron begins to *repolarize* itself, potentially leading to a charge below -65 mV, termed *hyperpolarization*. The interval of repolarization is known as the *refractory period*, where the neuron is inhibited from firing, thus suppressing any distribution of signals. After hyperpolarization, the neuron will slowly return to the resting state of -65 mV, thereby ending the whole firing sequence. Figure 1.2 displays this sequence of information processing within a neuron. The numbers point to the events outlined above: 1) the neuron receive stimulus input, 2) the neuron fires when the threshold is reached, 3) after firing the neuron might become hyperpolarized, 4) after repolarization, the neuron is once again at it's resting charge. If the membrane potentials threshold is

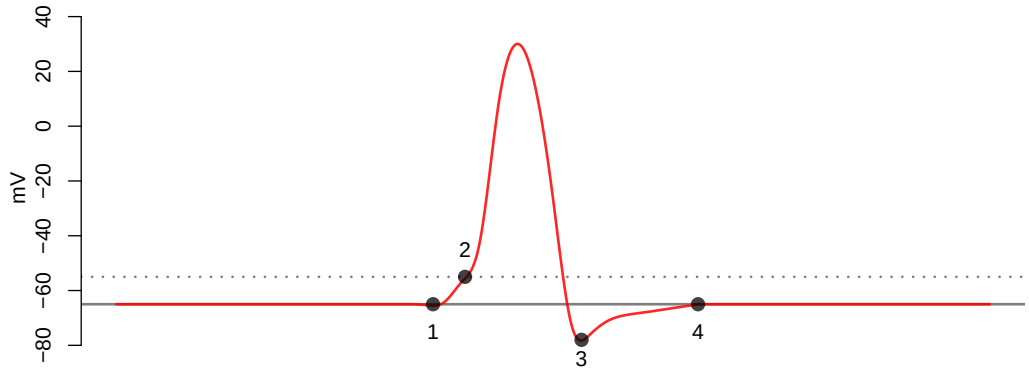

FIGURE 1.2: Propagation of an electrical impulse in a neuron. 1) Stimulus (signal) is received. 2) Threshold is reached and the neuron fires. 3) Neuron is hyperpolarized. 4) Neuron is repolarized.

not reached, the neuron will not fire and *subthreshold* oscillations may occur.

1.2.3 Spike Trains

When recording neurons, the full action potential, and thus subthreshold behavior, is available when recordings are *intra-cellular*. This is opposed to *extra-cellular* recordings where only the *spike times*, the time of firing, is available. Since the action potential has a positive width, it crosses the threshold twice, up and down. For this reason, spike times are typically defined as the upcrossing of the threshold, which would correspond to time 2) in Figure 1.2. Since the width of an action potential is very narrow, this definition does not lead to practical issues, but it is nonetheless important for a well defined mathematical description. A spike train(s) is then a sequence of recorded spike times from the same neuron(s). In Chapter 3 we introduce various representations of this type of data.

1.2.4 Neurons are Dynamical Systems

Conductance based mathematical models of neurons describe a neuron as a *dynamical system*, which implies that they model the evolution of the membrane potential, along with auxiliary variables, in the temporal dimension. The actual equations are then given as a set of differential equations, one for each variable/dimension of the model, with the most important variable being the membrane potential described above. Depending on whether a neuron is recorded intra-cellularly or extra-cellularly, we may or may not have a recording of the actual action potential. Since it is easier to record extra-cellularly, spike trains are often the only available data for analysis. However, conductance based models describe the complete shape of the action potential, replicating observations from intra-cellular recordings. Needless to say, this can obviously lead to identification issues, if multiple parameter settings produce similar spike times, when this is the only available information.

In Figure 1.2 we can interpret the mechanism initiating the firing, as integrating information. As the ion concentration become more positive, due to multiple inputs received at various dendrites, the neuron suddenly fires when reaching it's threshold. This understanding has lead to a simple class of one-dimensional neuron models termed *Integrate-and-Fire*, first presented by Lapicque (1907) (see Brunel and Rossum, 2007). A well-known version of this model type is the *leaky* integrate-and-fire model, where the membrane potential is described by a single differential equation on the temporal change in voltage (V) scaled by the constant capacitance of the neuron (C), controlled by the input current (I) and a function of the voltage and a voltage leak (E_{leak}), (see Izhikevich, 2010)

$$C\frac{dV}{dt} = I - g(V - E_{\text{leak}}). \tag{1.1}$$

This model is a simple construction, where the neuron fires an action potential once $V = E_{\text{threshold}}$ and instantly resets to a value E_K. Due to it's simplicity, (1.1) is not very accurate in replicating an observed action potentials, as it is simply modeling the spike time and not the shape of the action potential. Therefore more advanced (higher dimensional) models exist, some well known examples are Hodgkin and Huxley (1952), FitzHugh (1961), Nagumo, Arimoto, and Yoshizawa (1962), and Izhikevich (2003). All of these aim to improve the shortcomings of the integrate-and-fire model in various ways. However, common to all conductance based models is the modeling of the membrane potential V, through a temporal differential equation dV/dt, with added equations describing, excitatory-, inhibitory-, adaptive- and refractoriness features of the neuron. The Hodgkin-Huxley model is still considered one of the most physiologically realistic and meaningful models (Izhikevich, 2004), but it is very expensive computationally and parameter estimation is a cumbersome process. Regardless, it is considered a pinnacle of neuroscience models and the inventors, Alan Lloyd Hodgkin and Andrew Fielding Huxley received the Nobel prize in 1963 for their work and new models are often benchmarked to the Hodgkin-Huxley model or a simplification of it.

1.2.5 The Phase of a Neuron

Membrane potentials, and therefore action potentials, are most commonly described in terms of time and voltage, as in Figure 1.2. However, we can change focus and

instead look at the *phase* of the neuron. The waveform of the action potential has a beginning and an end. As such, we can interpret the membrane potential in this interval as rotation where the time of reset is 0 and the next spike/reset is at time T. By normalizing the interval such that the length becomes 2π, we can interpret the angle θ as the phase of the neuron. Figure 1.3 displays this relation where the

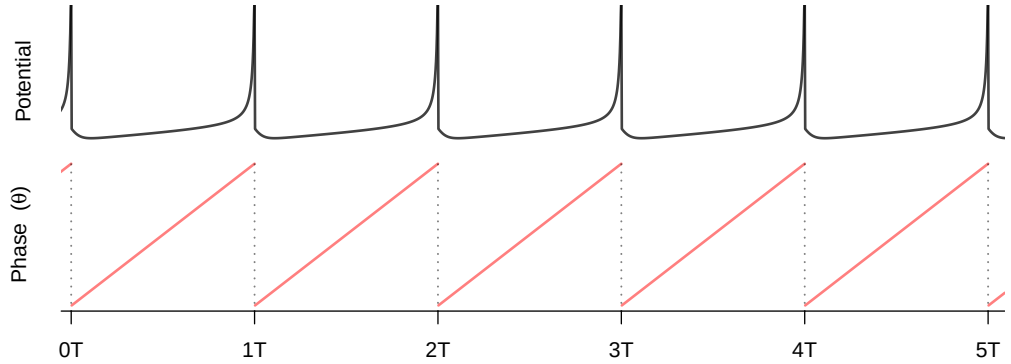

FIGURE 1.3: Membrane potential with spike times $0T, 1T, 2T, \ldots$. The corresponding phase below shows the relationship between the membrane potential and the angular interpretation of this.

membrane potential (top) is related to the phase (bottom). Each time a spike occurs, the potential is reset and so is the phase. It then grows linearly until the next spike, thus completing one revolution at T (or 2π). If we *unwrap* the phase process in Figure 1.3 we obtain a straight monotonically increasing line. The monotonicity is due to the regular spike intervals in the membrane potential, but if we include noise in the model, the corresponding phase process would start to wiggle, though overall it would still have a linear appearance, depending of course on the amount of noise added to the system. The type of continued spiking behavior in Figure 1.3 is termed *tonic spiking*. When a neuron exhibits continued spiking, we also refer to it as an *oscillator*, due to the repetitive spiking at (more or less) regular intervals with a monotonically (in the noiseless case) increasing phase process. In this case, modeling the phase process rather than the membrane potential, makes sense, since we are interested in modeling the spike times which depend directly on the phase: $\theta = 0$ or 2π when spiking.

1.2.6 Synchronization

When observing more than one oscillator, a certain phenomenon can present itself. This happens when the oscillators begin to adapt to each other, thus *synchronizing* their spike times. This behavior was first described by the Dutch mathematician Christiaan Huygens in the seventeenth century, when he discovered this peculiarity in two suspended pendulums. It is evident, that this event cannot appear unless there exist some form of *coupling* between the oscillators. A classic experiment displays this with a large number of metronomes on a platform. Each metronome is started, independently of the others with a certain beat. If the platform is fixed, no movement may travel between the metronomes and they never synchronize. However, if the platform is allowed to vibrate, even slightly, the metronomes will start to synchronize and after a while, all of them will beat in unison. A similar effect can take place in the theater when the audience applause. Initially, the applause will sound like random noise, every member of the audience is an individual oscillator,

but after a while one can observe a steadfast beat in clapping. Hence, the audience have synchronized their clapping. The phenomenon have also been observed in numerous places in nature, such as when fireflies synchronize their emission of bright flashes, or in physiological systems when neurons synchronize around a task. Chapter I presents a new approach to modeling coupled oscillators that adapt to each other, using cointegration analysis.

1.3 Previous Work

This thesis does not present the first attempt at bringing cointegration into the realm of neuroscience. Cointegration has been extensively studied in the econometrics community and major advancements have been with economic applications in mind. However, the work on cointegration in this thesis was inspired partially by the working paper by Dahlhaus and Neddermeyer (2012). This work made a strong attempt in analyzing a coupled chaotic Rössler/Lorenz system with an interpretation toward the phase process mentioned above. The paper linearly approximated the nonlinear Kuramoto (1984) system, which is only valid when the oscillators are closely phase-coupled, and thus, does not solve a more general system of oscillators. Dahlhaus and Neddermeyer (2012) also references other previous works that has attempted this bridging, without full success, but the ideas brought to light in this thesis have mainly been sparked by Dahlhaus and Neddermeyer (2012).

In regard to the second part of this thesis, generalized linear models are not new to neuroscience, but the idea of statistical models, as opposed to the conductance based models introduced above, is relatively new. Previous work has sought to describe the use of stochastic processes and the generalized linear model framework in the context of spike train analysis (Kass and Ventura, 2001; Truccolo et al., 2005). Newer work have looked into the generalized linear model frameworks capability in modeling various spike behaviors (Weber and Pillow, 2017), but where the generating model was without noisy input. The work on generalized models in this thesis attempts to demonstrate other aspects of this model class in terms of capturing variability in spike train data.

1.4 Overview

This thesis consist of two main parts. The first addresses the question of using cointegration analysis as a novel approach to analyze coupled networks of oscillators and detect the network structure. The second concentrate on analyzing spike trains (point processes) using generalized linear models and extensions of these. Each part open with an introductory chapter to establish a theoretical background for the included manuscripts.

1.4.1 Contributions

The contributions of this thesis consist of 4 manuscripts, 2 in each part. Two of these have been published in peer reviewed journals.

» Part 1 contain a published manuscript, *Oscillating systems with cointegrated phase processes*, on inferring the network coupling structure of a multivariate system

of oscillators. This paper describes the theoretical background for classical, discrete time, cointegration and the setting in continuous time. Then it is presented how a system of linearly coupled phase oscillators can be interpreted as a continuous time cointegrated system and it is demonstrated how to apply cointegration analysis to a coupled multivariate system. Finally a practical application is presented, where the technique is applied to EEG data for epileptic patients.

» The second paper of Part 1 is a working paper titled *Cointegration analysis of High-Dimensional linear Kuramoto networks*. This project explores cointegration analysis in a high-dimensional setting. Where as cointegration, historically, have been applied to systems up to dimension 10, in this paper we look at a stylized system: a linear version of the Kuramoto (1984) model, of coupled clusters and deal with rank testing in high dimensions and ways of detecting the inherent cluster structure.

» The first paper of Part 2 is the published work: *Capturing spike variability in noisy Izhikevich neurons using point process Generalized Linear Models*. This focuses on the use of generalized linear models as a statistical framework for analyzing point processes, more precisely spike trains. The purpose here is to demonstrate how these models capture essential features encoded in spike trains, by including the spike history, and performing model control by goodness-of-fit assessments. The paper addresses some of the shortcomings of the generalized linear model framework and discusses some possible extensions to overcome these.

» The second paper of Part 2 is titled *A state space model for bursting neurons*. It is a working paper which specifically deals with one of the shortcomings discussed in the previous manuscript, namely the bimodal distribution of intervals between spike times for bursting neurons, i.e. neurons that exhibit rapid firing (bursting) in a short period and then behave quiescent for a while before returning to a bursting sequence. The generalized linear model is extended with a latent state controlling this behavior and a point process version of the marginalized particle filter is applied for parameter estimation and determining latent states, simultaneously.

1.4.2 Further ideas

The work in this thesis has also inspired questions for future exploration. These include extensions to the current framework of cointegration to account for non-linear components and sparsity constraints. Another direction is the investigation of bootstrapping methods for misspecified models. Finally a completely new direction is the idea of cointegrated generalized linear models. A more specific presentation of these ideas is presented in Chapter 4.

Cointegration & Phase Synchronization

CHAPTER 2

Cointegration

2.1 Introduction

The first part of this thesis present cointegration analysis in a neuroscience context as a novel approach to identifying connections in neural networks. This part contributes with two manuscripts

- Oscillating systems with cointegrated phase processes (Østergaard, Rahbek, and Ditlevsen, 2017b)
- High-Dimensional cointegration in neural networks. (Østergaard, Rahbek, and Ditlevsen, 2017a)

The first paper, published in Journal of Mathematical Biology (Østergaard, Rahbek, and Ditlevsen, 2017b), presents a new approach to modeling connectivity of neurons and applies continuous time cointegration analysis to the phase processes of coupled oscillators. The paper defines the model class of linearly phase coupled oscillators as a multivariate stochastic differential equation and presents simulations based on various specifications from this class in the analyses. The paper also reviews important concepts of cointegration and places them in a continuous time setting, relevant to phase processes of neurons. Then it is demonstrated how a simulated network of coupled continuous oscillating processes can be analyzed using cointegration techniques. These oscillating processes are interpreted in a neuroscience context as neural processes from a multivariate system of connected neurons and it is shown that both independent, bi-directional as well as uni-directional couplings can be inferred. The paper also includes a short analysis on EEG data to illustrate the potential of cointegration, as a tool for neural data analysis. The research idea of the paper was to point out similarities between synchronized oscillators and the interpretation of cointegration for multivariate stochastic processes. In a synchronization context, the processes adapt to each other based on some common feedback within the system, referred to as trends. In a cointegration setting, the same can be stated. Here processes evolve individually with random walk like properties, but taken as a whole, the system exhibits common trends that are in fact stationary. The idea was to establish this connection by explaining the intuition behind cointegration and, by interpreting the constructed simulated processes in a neuroscience setting, thus demonstrating the capability of cointegration analysis in recovering the underlying network structure.

The second manuscript, which is still a working paper at this stage, adresses the fact that neural networks are often of a much higher dimension than currently tractable for cointegration analysis. In the classic setting, cointegration has been used with succes up to about 10 dimensions, but newer research studies are attempting to uncover feasible asymptotical distributions and estimation methods for a high-dimensional setting. Given the contemporary interest in high-dimensional statistics, this is an area of interest, not only for larger networks of connected neurons, but also in the area of econometrics where cointegration analysis is more common. As such, this paper pursues the quest of exploring the limits of classic cointegration techniques set in a high-dimensional context.

Both of these papers have been written in collaboration with Susanne Ditlevsen and Anders Rahbek, as projects in the research group *Dynamical Systems Interdisciplinary Network* (DSIN) at the University of Copenhagen.

2.2 Cointegration in Discrete Time

We begin by introducing classical results on cointegration in discrete time, including integrated processes, maximum likelihood estimation and inference. Then expanding to newer theory on bootstrap tests and continuous time models. The following section is based on Johansen (1996) and Lütkepohl (2005).

We start by introducing a multivariate version of a discrete time autoregressive (AR) stochastic process termed the *vector autoregressive* (VAR) process and a criterion for stationarity of such processes. Consider the process $y_n \in \mathbb{R}^p$

$$y_n = \sum_{i=1}^{k} \Pi_i y_{n-i} + \mu + \varepsilon_n, \text{ for } n \in \mathbb{Z} \tag{2.1}$$

where $\Pi_i \in \mathbb{R}^{p\times p}$ and ε_n is a p-dimensional white noise process such that $\mathbb{E}[\varepsilon_n] = 0$, $\mathbb{E}[\varepsilon_n \varepsilon_n'] = \Omega$ and $\mathbb{E}[\varepsilon_n \varepsilon_m'] = 0$ for $n \neq m$. We also assume that the ε_n process is Gaussian, i.e. $\varepsilon_n \sim \mathcal{N}_p(0, \Omega)$. The term μ is a fixed p-dimensional vector, thus implying the possibility of a non-zero mean of the process y_n.

For a VAR(1) process, we have that

$$\begin{aligned}
y_n &= \Pi_1 y_{n-1} + \mu + \varepsilon_n \\
&= \Pi_1(\Pi_1 y_{n-2} + \mu + \varepsilon_{n-1}) + \mu + \varepsilon_n \\
&= \Pi_1^2 y_{n-2} + (I_p + \Pi_1)\mu + \sum_{i=0}^{1} \Pi_1^i \varepsilon_{n-i} \\
&\vdots \\
&= \Pi_1^{j+1} y_{n-j-1} + \sum_{i=0}^{j} \Pi_1^i \mu + \sum_{i=0}^{j} \Pi_1^i \varepsilon_{n-i}.
\end{aligned}$$

This implies that if all eigenvalues of Π_1 have modulus less than 1, then $\lim_{j\to\infty} \Pi_1^j = 0$ and the following limit exist

$$\lim_{j\to\infty} \sum_{i=0}^{j} \Pi_1^i \mu = (I_p - \Pi_1)^{-1}\mu.$$

Furthermore, $\sum_{i=0}^{\infty} \Pi_1^i \varepsilon_{n-i}$ exists in L^2 and combining these statements we find that

$$y_n = (I_p - \Pi_1)^{-1}\mu + \sum_{i=0}^{\infty} \Pi_1^i \varepsilon_{n-i}, \text{ for } n \in \mathbb{Z}, \tag{2.2}$$

i.e. y_n can be represented as a vector moving average process. The eigenvalue criteria can be stated equivalently using the *characteristic polynomial*

$$\begin{gathered} \text{All eigenvalues of } \Pi_1 \text{ have modulus less than } 1 \\ \Updownarrow \\ |I_p - \Pi_1 z| \neq 0 \text{ for } |z| \leq 1, z \in \mathbb{C}, \end{gathered} \tag{2.3}$$

that is, the (complex) roots of the characteristic polynomial are outside the (complex) unit circle. A process satisfying (2.3) is said to be *stable*. Since a VAR(k) process can be represented as a VAR(1) process by defining

$$\begin{aligned} \mathbf{Y_n} &= (y_n, y_{n-1}, \ldots, y_{n-p+1})' \\ \boldsymbol{\mu} &= (\mu, 0, \ldots, 0)' \\ \mathbf{\Pi} &= \begin{pmatrix} \Pi_1 & \Pi_2 & \ldots & \Pi_{p-1} & \Pi_p \\ I_p & 0 & \ldots & 0 & 0 \\ 0 & I_p & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & I_p & 0 \end{pmatrix} \\ \boldsymbol{\varepsilon_n} &= (\varepsilon_n, 0, \ldots, 0)', \end{aligned}$$

we can use the arguments above to state that if

$$|A(z)| = |I_p - \mathbf{\Pi} z| = |I_p - \Pi_1 z - \Pi_2 z^2 - \cdots - \Pi_k z^k| \neq 0 \text{ for } |z| \leq 1, z \in \mathbb{C}, \tag{2.4}$$

the VAR(k) process y_n is stable. Since a stable VAR(k) is asymptotically stationary we also refer to y_n as asymptotically stationary, if the k-order VAR process satisfies (2.4). An asymptotically stationary process is strictly speaking not stationary, but the first and second order moments approach limit values as $n \to \infty$. For a stable, and thus asymptotically stationary, process without deterministic terms, the initial value can be given a distribution such that the process becomes stationary. For this reason the term "asymptotically" is often omitted in the literature on cointegration and we shall do so here as well. We therefore refer to a stable process as stationary, while in fact referring to asymptotically stationary with finite first and second order moments (also known as wide sense stationarity). Furthermore, we assume that a process is started at some time $n = 0$, such that time is positive $n = 0, 1, 2, \ldots$. This does not affect the results above, other than for the VAR(k) process we must include initial values $y_{-1}, \ldots, y_{-k+1}$ to define y_n in (2.1).

2.2.1 Integrated Processes

Assume the univariate random walk

$$x_n = x_{n-1} + \varepsilon_n, \text{ for } n \in \mathbb{Z} \tag{2.5}$$

with $\varepsilon_n \sim \mathcal{N}(0, \sigma^2)$. The condition (2.4) is not satisfied by this process as the characteristic polynomial

$$1 - z = 0 \tag{2.6}$$

clearly has a root at $z = 1$. Such a root for the characteristic polynomial is referred to as a *unit root* and we say that x_n is a unit root process. If the process starts at $n = 0$ with value x_0, then

$$x_n = x_0 + \sum_{i=1}^{n} \varepsilon_i$$

and we find the mean and variance of x_n

$$\begin{aligned} \mathbb{E}[x_n] &= x_0 \\ \text{Var}(x_n) &= \text{Var}(\sum_{i=1}^{n} \varepsilon_i) = n\sigma^2 \underset{n\to\infty}{\longrightarrow} \infty \end{aligned}$$

and the autocorrelation

$$\text{Cor}(x_n, x_{n+h}) = \frac{n}{\sqrt{n^2 + nh}} \underset{n\to\infty}{\longrightarrow} 1.$$

Hence, x_n and x_{n+h} will be highly correlated as n grows. This type of trending behavior is termed *stochastic trending*, since the trend of the process x_n is the cumulated random variables ε_n. If we include a constant term μ in (2.5) this becomes a random walk with drift, where the constant enters in the mean of the process as $\mathbb{E}[x_n] = x_0 + n\mu$, but the process is still stochastically trending with $\sum_{i=1}^{n} \varepsilon_i$. However, if we look at the differenced process (excluding the constant drift μ), this is stationary, since

$$\Delta x_n = x_n - x_{n-1} = \varepsilon_n \tag{2.7}$$

with

$$\begin{aligned} \mathbb{E}[\Delta x_n] &= 0 \\ \text{Var}(\Delta x_n) &= \sigma^2. \end{aligned}$$

For this reason, x_n is called *integrated of order 1*, denoted $I(1)$. For a general univariate AR(k) process, the characteristic polynomial can have multiple unit roots. For a unit root of order d, the d order differenced process

$$\Delta^d x_n = \Delta(\Delta^{d-1} x_n) = \Delta^{d-1} x_n - \Delta^{d-1} x_{n-1} = \varepsilon_n$$

is stationary and therefore such a process is referred to as integrated of order d, denoted $I(d)$. With these definitions, it makes sense to denote a stationary process $I(0)$. In this thesis we only consider processes that are $I(1)$, hence when referring

to an integrated process, it is implicitly understood that the process is $I(1)$ and thus have a stochastic trend similar to the random walk (2.5).

Extending the concept of integrated processes to multivariate processes is straightforward and we will refer to multivariate integrated processes as above, in the following text. Note that processes where $|z| = 1, z \in \mathbb{C}$ other than $z = 1$, corresponds to seasonal stochastic trends. This type of trend is out of the scope of this thesis, but see Johansen and Schaumburg (1999) for details on likelihood analysis for seasonal stochastic trends. Furthermore, if there are roots of the characteristic polynomial inside the complex unit circle the variance will tend to infinity at an exponential rate and hence the process is said to be explosive. Such processes are also not treated in this thesis and we will not discuss these concepts further.

2.2.2 Cointegrated VAR models

For a p-dimensional VAR(k) process y_n as in (2.1), where $|A(z)| = 0$ for $z = 1$ such that the process has a unit root and thus is $I(1)$, we can rewrite the VAR model as a differenced, and therefore $I(0)$, process

$$\begin{aligned}
\Delta y_n &= y_n - y_{n-1} \\
&= \Pi_1 y_{n-1} + \cdots + \Pi_k y_{n-k} + \varepsilon_n - y_{n-1} \\
&= -(I_p - \Pi_1) y_{n-1} + \Pi_2 y_{n-2} + \cdots + \Pi_k y_{n-k} + \varepsilon_n \\
&= -(I_p - \Pi_1 - \Pi_2) y_{n-1} - \Pi_2 y_{n-1} + \Pi_2 y_{n-2} + \cdots + \Pi_k y_{n-k} + \varepsilon_n \\
&= -(I_p - \Pi_1 - \Pi_2) y_{n-1} - \Pi_2 \Delta y_{n-1} + \Pi_3 y_{n-3} + \cdots + \Pi_k y_{n-k} + \varepsilon_n \\
&= -(I_p - \Pi_1 - \Pi_2 - \Pi_3) y_{n-1} - \Pi_3 y_{n-1} \\
&\quad - \Pi_2 \Delta y_{n-1} + \Pi_3 y_{n-2} - \Pi_3 y_{n-2} + \Pi_3 y_{n-3} + \cdots + \Pi_k y_{n-k} + \varepsilon_n \\
&= -(I_p - \Pi_1 - \Pi_2 - \Pi_3) y_{n-1} - (\Pi_2 + \Pi_3) \Delta y_{n-1} \\
&\quad - \Pi_3 \Delta y_{n-2} + \Pi_4 y_{n-4} + \cdots + \Pi_k y_{n-k} + \varepsilon_n \\
&\vdots \\
&= -(I_p - \Pi_1 - \cdots - \Pi_k) y_{n-1} - (\Pi_2 + \cdots + \Pi_k) \Delta y_{n-1} \\
&\quad - (\Pi_3 + \cdots + \Pi_k) \Delta y_{n-2} + \cdots - \Pi_k \Delta y_{n-k+1} + \varepsilon_n.
\end{aligned}$$

Collecting terms we find that

$$\Delta y_n = \Pi y_{n-1} + \sum_{i=1}^{k-1} \Gamma_i \Delta y_{n-i} + \varepsilon_n, \text{ for } n = 0, 1, 2, \ldots \tag{2.8}$$

where

$$\begin{aligned}
\Pi &= -(I_p - \Pi_1 - \cdots - \Pi_k), \\
\Gamma_j &= -(\Pi_{j+1} + \cdots + \Pi_k), \quad j = 1, \ldots, k-1.
\end{aligned}$$

Equation (2.8) is known as the *vector error correction model* (VECM) or the error correction form, of the VAR(k) process (2.1). Note here, that since y_n is $I(1)$, then Δy_n is $I(0)$ and since both sides of (2.8) are $I(0)$, the individual terms on the right hand side of (2.8) must all be $I(0)$. This especially implies that the term Πy_{n-1} is $I(0)$, but since y_n itself is $I(1)$, then the linear transformation Π must be a transformation of

the integrated process, such that it becomes $I(0)$, i.e., stationary. For the k-order VAR process above, we have that

$$|A(z)| = |I_p - \Pi_1 z - \Pi_2 z^2 - \cdots - \Pi_k z^k| = 0, \text{ for } z = 1,$$

implying that

$$|\Pi| = -|A(1)| = -|I_p - \Pi_1 - \Pi_2 - \cdots - \Pi_k| = 0. \tag{2.9}$$

Thus, $\Pi \in \mathbb{R}^{p\times p}$ is singular and therefore not of full rank $r < p$. In general, we can write three exhaustive possibilities for the rank of a matrix $M \in \mathbb{R}^{p\times p}$, either

1. $\text{rank}(M) = 0 \Rightarrow M = 0$, or
2. $\text{rank}(M) = p \Rightarrow M$ is non-singular and hence invertible, or
3. $\text{rank}(M) \in (0, p) \Rightarrow M$ is singular, i.e., non-invertible.

If $\Pi = 0$, then (2.8) is a stationary VAR(k) for $x_n = \Delta y_n$. If Π is non-singular, then it has full rank and $|A(z)| \neq 0$ for $z = 1$, meaning that the one-to-one transformation Πy_n is $I(0)$ and hence $\Pi^{-1}\Pi y_n = y_n$ is $I(0)$. However, if we consider the singular case, then $\text{rank}(\Pi) = r$ with $0 < r < p$ and $|\Pi| = 0$, and as such we are in the regime of integrated processes with stationary linear combinations, also known as *cointegrated processes*.

Assume therefore, that Π is singular with $\text{rank}(\Pi) = r < p$. We can then write Π as the product of two matrices $\alpha, \beta \in \mathbb{R}^{p\times r}$ of full rank r

$$\Pi = \alpha\beta'. \tag{2.10}$$

Since $\alpha'\alpha \in \mathbb{R}^{r\times r}$ has full rank r, the inverse exist, and from the discussion above, order of integration is preserved when performing a non-singular transformation, hence

$$(\alpha'\alpha)^{-1}\alpha'\Pi y_{n-1} = (\alpha'\alpha)^{-1}\alpha'\alpha\beta' y_{n-1} = \beta' y_{n-1}$$

is $I(0)$, or in other words, the r-dimensional column space of β determine the subspace $\text{sp}(\beta) \subset \mathbb{R}^p$ containing the cointegration relations. For this reason, β is sometimes termed the *cointegration matrix* and α the *loading matrix*. It is important to note that the decomposition $\alpha\beta'$ is not unique, since we can simply multiply either matrix with a non-singular matrix $M \in \mathbb{R}^{r\times r}$ and obtain another stationary transformation $\tilde{\Pi} y_n$ of y_n.

In order to present a fundamental result in cointegration theory, we define the orthogonal complement to a full rank matrix $M \in \mathbb{R}^{p\times r}$, $r \leq p$ as $M_\perp \in \mathbb{R}^{p\times(p-r)}$, such that $M_\perp$ has rank $p - r$ and $M'M_\perp = 0$. If $r = p$ we define $M_\perp = 0$ and if $r = 0$ then $M_\perp = I_p$. With this definition we can state an important theorem on cointegrated VAR models, known as the Granger Representation Theorem.

Granger Representation Theorem (Engle and Granger, 1987)

Given the p-dimensional process y_n on error correction form (2.8) and assuming that

- $|A(z)| = 0$ *implies that* $|z| > 1$ *or* $z = 1$
- $\text{rank}(\Pi) = r < p$

- $\Pi = \alpha\beta'$ *for matrices* $\alpha, \beta \in \mathbb{R}^{p\times r}$ *of rank* r*,*

then a necessary and sufficient condition, such that $\Delta y_n - \mathbb{E}[\Delta y_n]$ *and* $\beta' y_n - \mathbb{E}[\beta' y_n]$ *can be given initial distributions so that they are* $I(0)$ *is that*

$$|\alpha'_\perp \Gamma \beta_\perp| \neq 0, \tag{2.11}$$

where $\Gamma = I_p - \sum_{i=1}^{k-1} \Gamma_i$*. Furthermore, if (2.11) is true, then*

$$y_n = C\sum_{i=1}^{n} \varepsilon_n + C_1(L)\varepsilon_n + y_0^*, \tag{2.12}$$

where

- $C = \beta_\perp(\alpha'_\perp \Gamma \beta_\perp)^{-1}\alpha'_\perp$*,*
- $C_1(z)$ *is a matrix polynomial satisfying*

$$A^{-1}(z) = C\frac{1}{1-z} + C_1(z), \text{ for } z \neq 1,$$

- y_0^* *contains initial values such that* $\beta' y_0^* = 0$*.*

In the above, L denotes the lag operator such that $Ly_n = y_{n-1}$. For a proof of the theorem, see Johansen (1996).

Equation (2.12) splits the process y_n into $I(0)$ and $I(1)$ components, since

$$C\sum_{i=1}^{n} \varepsilon_n \text{ is } I(1)$$
$$C_1(L)\varepsilon_n \text{ is } I(0),$$

and since $\beta' y_0^* = \beta' C = 0$, then $\beta' y_n$ is $I(0)$. Thus, even though y_n is $I(1)$ and a pure moving average representation does not exist, then under the assumptions of the Granger Representation Theorem we can decompose the solution to (2.8) into a part related to stationary trends and a part related to stochastic trends. The matrix C has rank $p - r$ and explains how the common stochastic trends $\alpha'_\perp \sum_{i=1}^{n} \varepsilon_i$ affect the system through the matrix $\beta_\perp$. Furthermore, the so callled *long run variance* of y_n is given by the singular composition $C\Omega C'$.

As a consequence of the discussion above we find that, whereas the concept of integrated processes applies to both univariate and multivariate processes, cointegration is a concept present only in multivariate settings. The decomposition in the Granger Representation Theorem especially tells us, that for a p-dimensional cointegrated process, with $\text{rank}(\Pi) = r$, then there are exactly r cointegrating relations, or stationary trends and $p - r$ common stochastic (random walk type) trends. The discussion related to cointegrated processes can be extended to include deterministic terms in the error correction form (2.8) or the VAR form (2.1). The Granger Representation is then slightly modified to account for these terms. In the above, we have omitted these terms for added clarity, but for a complete discussion on the more general theorem including deterministic terms, see Johansen (1996).

In the bivariate setting we can visualize cointegration as a process revolving around an attractor, given by the span of $\beta_\perp$. This is shown in Figure 2.1 for the system

$y_n = (y_n^{(1)}, y_n^{(2)})'$, where the common stochastic trends $\alpha'_\perp \sum_{i=1} \varepsilon_n$ push the process along the attractor set sp$(\beta_\perp)$, while the cointegration relations pull the process to-

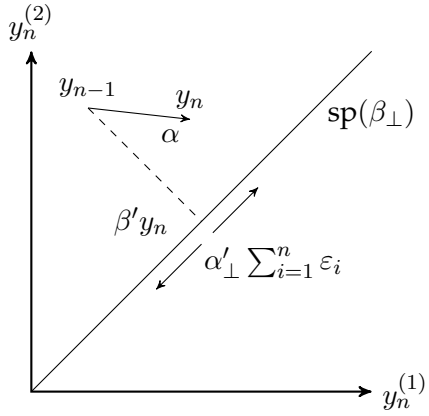

FIGURE 2.1: The process y_n is pushed towards the attractor sp$(\beta_\perp)$, while the common stochastic trends $\alpha'_\perp \sum_{i=1}^n \varepsilon_i$ pushes along the attractor. Reproduced from Johansen (1996).

wards the attractor. Note that $\beta' y_n$ corresponds to the projection of y_n onto the attractor set. Another intuitive interpretation is that the relation $\beta' y_n$ can be viewed as a kind of long run "equilibrium" relation that the process y_n will adhere to, but it is not something that is eventually reached, such that the process stays in equilibrium after a certain period. Instead we can think of it as a rubber constraint. When the disequilibrium is larger, the process will experience a stronger pull towards sp(β), whereas for smaller disequilibriums the process is not as strongly pulled towards the attractor. Also, in regard to the stochasticity, even minor shocks to parts of the system will eventually be propagated into the full system, but the total effect will die out as time progresses and the cumulative effect of the stochastic trend eventually dominates any short term randomness. Hence, the major effects on y_n can be summed up as the r $\beta' y_n$ trends and $p - r$ $\alpha'_\perp \sum_{i=1}^n \varepsilon_i$ trends.

2.2.3 Maximum Likelihood Estimation

We review maximum likelihood estimation for cointegrated VAR models with i.i.d Gaussian errors (white noise) using reduced rank regression. This is popularly referred to as the "Johansen Procedure" in the literature. While other estimation procedures exist, maximum likelihood estimation provides both parameter estimates and determines $\mathrm{rank}(\Pi)$ simultaneously. Because of this it is a popular method for cointegrated processes.

We will assume the VAR(k) process y_n and that the assumptions for the Granger Representation Theorem are satisfied, such that y_n is $I(1)$ with $\mathrm{rank}(\Pi) = r$. We omit the inclusion of deterministic trends in the model, but the generalization to include these is straightforward with respect to estimation (see Johansen, 1996). However, in order to determine the rank of Π, this is not the case. This task requires a precise model specification with any deterministic terms and the corresponding restrictions, if any, for these terms. This is necessary in order to obtain the correct asymptotic distribution for the rank test. But for now we simply assume that $\mathrm{rank}(\Pi) = r$ is

known and y_n is on the form (2.8) which we repeat here,

$$\Delta y_n = \alpha\beta' y_{n-1} + \sum_{i=1}^{k-1} \Gamma_i \Delta y_{n-i} + \varepsilon_n, \text{ for } n = 1, \ldots, N.$$

Note that the only difference with (2.8) is that here we have assumed a total of N observations from the process rather than $n \in \mathbb{N}_0$ as in (2.8) and explicitly decomposed $\Pi = \alpha\beta'$. Given initial values $y_{-k+1}, \ldots, y_0$, the parameters of the model are $(\alpha, \beta, \Gamma_1, \ldots, \Gamma_{k-1}, \Omega)$. Define now $z_{0n} = \Delta y_n$ and $z_{1n} = y_{n-1}$. Furthermore let $z_{2n} = (\Delta y_{n-1}, \ldots, \Delta y_{n-k+1})$ denote the $p(k-1)$-dimensional stacked vector of lagged variables. Finally let $\Theta = (\Gamma_1, \ldots, \Gamma_{k-1})'$ denote the corresponding $p \times p(k-1)$ matrix of parameters for z_{2n}. Then (2.8) becomes

$$z_{0n} = \alpha\beta' z_{1n} + \Theta z_{2n} + \varepsilon_n, \text{ for } n = 1, \ldots, N. \tag{2.13}$$

Since $\varepsilon_n \sim \mathcal{N}(0, \Omega)$ we can write the log-likelihood function of (2.13) as

$$\begin{aligned} &\log L(Y_n, \alpha, \beta, \Theta, \Omega) \\ &\quad = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log|\Omega| - \frac{1}{2}\sum_{n=1}^{N}(z_{0n} - \alpha\beta' z_{1n} - \Theta z_{2n})'\Omega^{-1}(z_{0n} - \alpha\beta' z_{1n} - \Theta z_{2n}). \end{aligned} \tag{2.14}$$

Differentiating with respect to Θ and equating to 0 yields the estimation equation for Θ

$$\sum_{n=1}^{N} z_{0n} z_{2n}' - \alpha\beta' z_{1n} z_{2n}' = \hat{\Theta} \sum_{n=1}^{N} z_{2n} z_{2n}'. \tag{2.15}$$

By introducing the notation for product moment matrices of $z_{in}, i = 0, 1, 2$

$$M_{ij} = N^{-1} \sum_{n=1}^{N} z_{in} z_{jn}', \text{ for } i, j = 0, 1, 2 \tag{2.16}$$

then for α, β fixed, we obtain the estimate of Θ

$$\hat{\Theta} = M_{02} M_{22}^{-1} - \alpha\beta' M_{12} M_{22}^{-1}.$$

Defining the residuals from regressing z_{0n} and z_{1n} on z_{2n} as

$$\begin{aligned} R_{0n} &= z_{0n} - M_{02} M_{22}^{-1} z_{2n} \\ R_{1n} &= z_{1n} - M_{12} M_{22}^{-1} z_{2n} \end{aligned}$$

and ignoring constant terms, we obtain the profile log-likelihood

$$\log L(\alpha, \beta, \Omega) = -\frac{N}{2}\log|\Omega| - \frac{1}{2}\sum_{n=1}^{N}(R_{0n} - \alpha\beta' R_{1n})'\Omega^{-1}(R_{0n} - \alpha\beta' R_{1n}). \tag{2.17}$$

corresponding to a reduced rank regression of R_{0n} on R_{1n} since $\alpha\beta'$ has rank $r < p$. Note that with this construction, we have effectively removed the influence of the lagged differences in z_{2n} and what remains is a regression problem concerning only the reduced matrix $\alpha\beta'$. For this we need to solve an eigenvalue problem to obtain

an estimate of β. Then given β, estimates $\hat{\alpha}$ and $\hat{\Omega}$ are obtained by regressing R_{0n} on $\beta' R_{1n}$. Define yet another set of moment matrices with respect to the residuals R_{0n} and R_{1n}

$$S_{ij} = N^{-1} \sum_{n=1}^{N} R_{in} R'_{jn},$$

then differentiating (2.17) with respect to α, equating to zero and solving we find

$$S_{01}\beta = N^{-1} \sum_{n=1}^{N} R_{0n} R'_{1n} \beta = N^{-1} \sum_{n=1}^{N} \hat{\alpha}\beta' R_{1n} R'_{1n} \beta = \hat{\alpha}\beta' S_{11} \beta$$

such that

$$\hat{\alpha} = S_{01}\beta(\beta' S_{11}\beta)^{-1}. \tag{2.18}$$

Similarly, for β fixed, an estimate of the covariance matrix is obtained as

$$\begin{aligned}
\hat{\Omega} &= N^{-1} \sum_{n=1}^{N} (R_{0n} - \hat{\alpha}\beta' R_{1n})(R_{0n} - \hat{\alpha}\beta' R_{1n})' \\
&= N^{-1} \sum_{n=1}^{N} R_{0n} R'_{0n} - \hat{\alpha}\beta' R_{1n} R'_{1n} \beta \hat{\alpha}' \\
&= S_{00} - \hat{\alpha}\beta' S_{11} \beta \hat{\alpha}' \\
&= S_{00} - S_{01}\beta(\beta' S_{11}\beta)^{-1} \beta' S_{10},
\end{aligned} \tag{2.19}$$

where we used that $S_{ij} = S'_{ji}$. The only remainder now, is the estimate of β. From estimation of α, Ω and Θ, the likelihood is proportional to

$$L(\hat{\alpha}, \beta, \hat{\Omega}, \hat{\Theta}) \propto |\hat{\Omega}|,$$

where $\hat{\Omega}$ depends on β. Writing $|\hat{\Omega}|$ by exploiting the Schur complement as

$$\begin{aligned}
|S_{00} - S_{01}\beta(\beta' S_{11}\beta)^{-1}\beta' S_{10}| &= |S_{00}||\beta' S_{11}\beta - \beta' S_{10} S_{00}^{-1} S_{01})\beta|/|\beta' S_{11}\beta| \\
&= |S_{00}||\beta'(S_{11} - S_{10} S_{00}^{-1} S_{01})\beta|/|\beta' S_{11}\beta|,
\end{aligned} \tag{2.20}$$

the likelihood is maximized with respect to β' by solving for eigenvalues λ of

$$|\lambda S_{11} - S_{10} S_{00}^{-1} S_{01}| = 0, \tag{2.21}$$

such that for eigenvalues λ_i and corresponding eigenvectors v_i

$$\lambda_i S_{11} v_i = S_{10} S_{00}^{-1} S_{01} v_i \text{ for } i = 1, \ldots, p, \tag{2.22}$$

where the eigenvectors are normalized with

$$v'_j S_{11} v_i = \begin{cases} 1 \text{ for } i = j \\ 0 \text{ for } i \neq j. \end{cases} \tag{2.23}$$

With the assumption of $\mathrm{rank}(\Pi) = r$, the estimated cointegration relations, i.e. the

column space of $\hat{\beta}$, is the space spanned by the eigenvectors $(v_1, \ldots, v_r)$ corresponding to the r largest eigenvalues: $\lambda_1 > \cdots > \lambda_r > \cdots > \lambda_p$. As noted above, the estimation of $\hat{\alpha}\hat{\beta}'$ is not unique. Hence, for identifiability of α and β it is common to impose a normalization on $\hat{\beta}$. Let $c = (I_r, 0_{p-r\times r})'$ denote a $p \times r$ matrix with the identity stacked on a zero matrix, then for any version $\tilde{\beta}$ of $v_1, \ldots, v_r$ we can use

$$\hat{\beta} = \tilde{\beta}(c'\tilde{\beta})^{-1} \tag{2.24}$$

as a normalizing factor such that

$$\hat{\beta} = \begin{pmatrix} I_r \\ \tilde{\beta}_{p-r\times r} \end{pmatrix}, \tag{2.25}$$

where $\tilde{\beta}_{p-r\times r}$ refers to the last $p - r$ rows of $\tilde{\beta}$.

2.2.4 Estimation of the cointegration rank

Having established the maximum likelihood estimation of parameters $\alpha, \beta, \Gamma_1, \ldots, \Gamma_{k-1}, \Omega$, we are left with the question of the rank of Π. Obviously this is a crucial initial step in cointegration analysis, since $\mathrm{rank}(\Pi)$ determines the number of cointegration trends and stochastic trends, i.e. $I(0)$ and $I(1)$ trends.

From (2.23) we have

$$v_j' S_{10} S_{00}^{-1} S_{01} v_i = \lambda_i v_j' S_{11} v_i = \begin{cases} \lambda_i \text{ for } i = j \\ 0 \text{ for } i \neq j. \end{cases}$$

Hence, the eigenvectors v_i simultaneously diagonalize $S_{10}S_{00}^{-1}S_{01}$ and S_{11}. This means that the likelihood, apart from a constant, is maximized as

$$\begin{aligned} L_{\max} &= |S_{00}| \frac{|\hat{\beta}'(S_{11} - S_{10}S_{00}^{-1}S_{01})\hat{\beta}|}{|\hat{\beta}' S_{11} \hat{\beta}|} \\ &= |S_{00}| \prod_{i=1}^{r} (1 - \hat{\lambda}_i), r = 0, \ldots, p. \end{aligned} \tag{2.26}$$

For $r = 0$ we set $\hat{\beta} = 0$, i.e. $\hat{\Pi} = 0$ and for $r = p$, $\hat{\beta} = I_p$ and hence $\hat{\Pi} = \hat{\alpha} = S_{01}S_{11}^{-1}$. As a result, (2.26) maximizes the likelihood function for all possible choices of r, since we have already solved for all p eigenvalues in (2.22).

To form a likelihood ratio test of r, first note that defining the hypothesis $H(r)$ as $\mathrm{rank}(\Pi) \leq r$, then the sequence of hypothesis for $r = 0, \ldots, p$ are nested as

$$H(0) \subset H(1) \subset \cdots \subset H(p), \tag{2.27}$$

since any model of rank r is included in a model of rank $r + 1$. The extremes, $H(0), H(p)$ corresponds to the restriction $\Pi = 0$ and an unrestricted VAR(k) model. That is a pure $I(1)$ model with p stochastic trends, versus a pure $I(0)$ model with no stochastic trends. Using the nested sequence (2.27) we can form a likelihood ratio

test of $H(r)$ versus $H(p)$

$$Q\big(H(r)|H(p)\big)^{-2/N} = \frac{|S_{00}|\prod_{i=1}^{r}(1-\hat{\lambda}_i)}{|S_{00}|\prod_{i=1}^{p}(1-\hat{\lambda}_i)}$$

which, expressed in log form, is equivalent to

$$-2\log Q\big(H(r)|H(p)\big) = -N\sum_{i=r+1}^{p}\log(1-\hat{\lambda}). \tag{2.28}$$

Testing for $H(r)$ versus $H(r+1)$ is also a possible strategy, by reformulating (2.28) as

$$-2\log Q\big(H(r)|H(r+1)\big) = -N\log(1-\hat{\lambda}_{r+1}). \tag{2.29}$$

The statistics (2.28) and (2.29) are referred to as the *trace statistic* and *maximum eigenvalue statistic* respectively. The names are based on the non-standard asymptotical distribution of the statistics under the null hypotheses,

$$\begin{aligned} -2\log Q\big(H(r)|H(p)\big) &\xrightarrow{d} \mathrm{Tr}(\mathcal{D}) \\ -2\log Q\big(H(r)|H(r+1)\big) &\xrightarrow{d} \lambda_{\max}(\mathcal{D}), \end{aligned}$$

where $\mathrm{Tr}(\mathcal{D})$ and $\lambda_{\max}(\mathcal{D})$ refer to the trace and the maximum eigenvalue, respectively, of the matrix

$$\mathcal{D} = \left(\int_0^1 F dW'\right)'\left(\int_0^1 FF'ds\right)^{-1}\left(\int_0^1 F dW'\right).$$

Here W denote a $p-r$-dimensional Wiener process and F depends on the form and restrictions of any deterministic terms included in the model, as mentioned in the beginning of this section. In the case here of no deterministic terms, then $F = W$. Due to the non-standard asymptotical distributions for (2.28) and (2.29), critical values can be found by simulation. Tabulated values for various combinations of dimension and model specifications are available in Johansen (1996).

Determining $\hat{r} = \mathrm{rank}(\Pi)$ using the trace statistic (2.28) is then performed as a sequential procedure starting from $H(0)$ versus $H(p)$ and the continuing until the null cannot be rejected. That is, starting with $r = 0$, do the following

1. Calculate $\Lambda(r) = -2\log Q\big(H(r)|H(p)\big)$ and evaluate.
2. If $\Lambda(r)$ is rejected, set $r = r+1$ and go to 1, otherwise set $\hat{r} = r$.
3. If $r = p$, then set $\hat{r} = p$.

2.2.5 Bootstrap testing for cointegration rank

Another possibility for evaluating the critical thresholds for the statistics (2.28) and (2.29), is to use bootstrapping techniques (see Cavaliere, Rahbek, and Taylor, 2012), which also improves the small sample properties of the test.

The intuition stems from the fact that the innovations in (2.8) are i.i.d. Gaussian. This implies that given estimates $\hat{\alpha}, \hat{\beta}, \hat{\Theta}, \hat{\Omega}$, we can simulate data from the estimated

model by simulating Gaussian innovations. Then for each such sample, we calculate the relevant statistic and evaluate our original statistic to the empirical distribution of bootstrap sampled values.

In Cavaliere, Rahbek, and Taylor (2012) the bootstrap algorithm for rank estimation of the model (2.8) is presented as follows. Given N observations, set $r = 0, \hat{\alpha} = \hat{\beta} = 0$ and do

1. Estimate the parameters of (2.8) by MLE as presented in the previous section. Denote the estimates $\hat{\alpha}^{(r)}, \hat{\beta}^{(r)}, \hat{\Theta}^{(r)}, \hat{\Omega}^{(r)}$.
2. Generate N bootstrap innovations, denoted ε_n^* from the residuals $\hat{\varepsilon}_n, n = 1, \ldots, N$ from (2.8) by either
   (a) *The i.i.d. bootstrap*: $\varepsilon_n^* = \hat{\varepsilon}_{u_n}, n = 1, \ldots, N$ for an i.i.d sequence of discrete uniform variables u_n on $(1, \ldots, N)$.
   (b) *The wild bootstrap*: $\varepsilon_n^* = \hat{\varepsilon}_n w_n, n = 1, \ldots, N$ for an i.i.d. sequence of variables $w_n \sim \mathcal{N}(0, 1)$.
3. Using the bootstrap innovations ε_n^* and initial values $y_{-k+1}^* = \cdots = y_0 = 0$, sample recursively from

$$\Delta y_n^* = \hat{\alpha}^{(r)} \hat{\beta}^{(r)} y_{n-1}^* + \sum_{i=1}^{k} \hat{\Gamma}_i^{(r)} \Delta y_{n-i}^* + \varepsilon_n^*, \text{ for } n = 1, \ldots, N.$$

4. Use the bootstrap sample $\{y_n^*\}_{n=1}^N$ to find the eigenvalues $\lambda_1^* > \cdots > \lambda_p^*$ as in (2.22) and construct either the trace statistic (2.28) or the maximum eigenvalue statistic (2.29) denoted here by Q_r^*.
5. Evaluate the p_r^*-value for the bootstrap sample with the (unknown) distribution G_r^*. If p_r^* exceeds the significance level, set $\hat{r} = r$, otherwise set r=r+1 and go to 1. If $r + 1 = p$ set $\hat{r} = p$.

The distribution G_r^* in step 5 is unknown, but it can be approximated from independent bootstrap samples. By repeating steps 1-4 above for B bootstrap samples and denoting each sampled statistic as $Q_r^{(b)}$ for $b = 1, \ldots, B$, we obtain an empirical distribution $\tilde{G}_r^*$ for the statistic, approximating G_r^*. We can then evaluate the original statistic Q_r, as calculated from the original data, with the empirical $\tilde{G}_r^*$. For the B bootstrap samples of $Q_r^{(b)}, b = 1, \ldots, B$, the p-value of Q_r is evaluated with $\tilde{G}_r^*$ as

$$\tilde{p}_r^* = B^{-1} \sum_{i=1}^{B} \mathbb{1}\{Q_r^{(b)} > Q_r\}. \tag{2.30}$$

The fact that $\tilde{p}_r^* \to p_r^*$ with asymptotical standard error $\sqrt{\tilde{p}_r^*(1 - \tilde{p}_r^*)/B} \to 0$ as $B \to \infty$, where p_r^* denotes the true p-value for Q_r evaluated by the distribution G_r^*, assures us that for a large enough bootstrap sample, B, the p-values from the empirical distribution are valid for rank estimation.

2.2.6 Inference for α and β

We now consider restrictions for the parameters α and β in (2.8). Recall that the estimate of $\hat{\Pi} = \hat{\alpha}\hat{\beta}'$ was not unique, since we are actually identifying the subspaces $\text{sp}(\alpha)$, $\text{sp}(\beta)$ and it was necessary to impose a normalization such as (2.24) on $\hat{\beta}$ to identify parameters.

These considerations also apply to the possible inference on the cointegration parameters. We can extend the intuition behind the normalization (2.24) to more general linear restrictions on either α or β. To keep things simple, we will illustrate examples for a $p = 3$ dimensional system of (possibly) coupled neurons, where $r = 1, 2$. Denote the process as $y_n = (y_{1n}, y_{2n}, y_{3n})'$ and likewise for Δy_n. For simplicity we also assume that no deterministic terms or lagged differences Δy_{n-i} are present. Hence, the model is simply

$$\Delta y_n = \begin{pmatrix} \Delta y_{1n} \\ \Delta y_{2n} \\ \Delta y_{3n} \end{pmatrix} = \alpha\beta' \begin{pmatrix} y_{1n} \\ y_{2n} \\ y_{3n} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1n} \\ \varepsilon_{2n} \\ \varepsilon_{3n} \end{pmatrix}. \tag{2.31}$$

The case of $r = 1$ implies that there is 1 cointegration relation, i.e., α, β are column vectors in $\mathbb{R}^3$, and two stochastic trends. If we have reason to assume that y_{1n} and y_{2n} are coupled independently of y_{3n}, we could think of this as stating that y_{1n}, y_{2n} cointegrate and share a common stochastic trend, whereas y_{3n} has it's own stochastic trend, independent of y_{1n}, y_{2n}. In a cointegration context, this would imply that

$$H_0: \quad \Pi = \begin{pmatrix} \pi_{11} & \pi_{12} & 0 \\ \pi_{21} & \pi_{22} & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{2.32}$$

If we allowed for y_{3n} to influence y_{1n} and y_{2n}, but not the other way around, then we would be interested in whether

$$H_0: \quad \Pi = \begin{pmatrix} \pi_{11} & \pi_{12} & \pi_{13} \\ \pi_{21} & \pi_{22} & \pi_{23} \\ 0 & 0 & 0 \end{pmatrix} \tag{2.33}$$

is a plausible explanation of the data. These hypothesis can be tested by restrictions on α and/or β on the form

$$\begin{matrix} R_\alpha'\alpha = 0 \\ R_\beta'\beta = 0 \end{matrix} \quad \Leftrightarrow \quad \begin{matrix} \alpha = H_\alpha\psi \\ \beta = H_\beta\phi \end{matrix} \tag{2.34}$$

for matrices $H = R_\perp$ of size $p \times s_i$ and ψ, ϕ are matrices of size $s_i \times r$ for $i = \alpha, \beta$. The hypothesis (2.33) would then be specified as

$$\begin{aligned} H_0: \quad & \alpha = H_\alpha\psi \\ & H_\alpha = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \\ & \psi = (\psi_1, \psi_2)' \end{aligned}$$

such that $\alpha = (\psi_1, \psi_2, 0)'$ and therefore

$$\Pi = \alpha\beta' = \begin{pmatrix} \psi_1\beta_1 & \psi_1\beta_2 & \psi_1\beta_3 \\ \psi_2\beta_1 & \psi_2\beta_2 & \psi_2\beta_3 \\ 0 & 0 & 0 \end{pmatrix}.$$

Note that the restriction in this case would be $R_\alpha = (0, 0, 1)'$. A similar restriction on β would amount to

$$\begin{aligned} H_0: \quad \beta &= H_\beta \phi \\ H_\beta &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \\ \phi &= (\phi_1, \phi_2)' \end{aligned}$$

and thus $\beta = (\phi_1, \phi_2, 0)'$

$$\Pi = \alpha\beta' = \begin{pmatrix} \alpha_1\phi_1 & \alpha_1\phi_2 & 0 \\ \alpha_2\phi_1 & \alpha_2\phi_2 & 0 \\ \alpha_3\phi_1 & \alpha_3\phi_2 & 0 \end{pmatrix},$$

with a restriction similar to the previous case, $R_\beta = (0, 0, 1)'$. Evident from the Π matrix above, the cointegration relation $\beta' y_n$ would ensure that y_{3n} does not influence either y_{1n} or y_{2n}, but if $\alpha_3 \neq 0$, then y_{3n} would be influenced by any disequilibrium between y_{1n} and y_{2n} and therefore *not* independent. Hence, in order to obtain full independence for y_{3n}, then simultaneous restrictions on α and β must be imposed

$$\begin{aligned} H_0: \quad \alpha = H_\alpha \psi \quad \wedge \quad \beta &= H_\beta \phi \\ H_\alpha &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \\ \psi &= (\psi_1, \psi_2)' \\ H_\beta &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \\ \phi &= (\phi_1, \phi_2)' \end{aligned}$$

implying that

$$\Pi = \alpha\beta' = \begin{pmatrix} \psi_1\phi_1 & \psi_1\phi_2 & 0 \\ \psi_2\phi_1 & \psi_2\phi_2 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

As a final example, assume that one process, say y_{1n} is a random walk that influences both y_{2n} and y_{3n}, such that

$$\Delta y_n = \begin{pmatrix} \Delta y_{1n} \\ \Delta y_{2n} \\ \Delta y_{3n} \end{pmatrix} = \begin{pmatrix} 0 \\ \alpha_2\beta_1 \\ \alpha_3\beta_1 \end{pmatrix} y_{1n} + \begin{pmatrix} \varepsilon_{1n} \\ \varepsilon_{2n} \\ \varepsilon_{3n} \end{pmatrix}. \tag{2.35}$$

For $r = 1$ this would have $\alpha = (0, \alpha_2, \alpha_3)'$ and $\beta = (\beta_1, 0, 0)'$, corresponding to the α restriction

$$\begin{aligned} H_0: \quad \alpha &= H_\alpha \psi \\ H_\alpha &= \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \\ \psi &= (\psi_2, \psi_3)'. \end{aligned}$$

Since $r = 1$ there are $p - r = 2$ common stochastic trends given by $\alpha_\perp' \sum_{i=1}^n \varepsilon_i$. From the form of α we find that

$$\alpha_\perp = \begin{pmatrix} \tilde{\alpha}_1 & \tilde{\alpha}_2 \\ -\alpha_3 & -\alpha_3 \\ \alpha_2 & \alpha_2 \end{pmatrix},$$

such that $\alpha_\perp' \alpha = 0$. This implies that

$$\begin{aligned} \alpha_\perp' \sum_{i=1}^n \varepsilon_n &= \begin{pmatrix} \tilde{\alpha}_1 \sum_{i=1}^n \varepsilon_{1n} - \alpha_3 \sum_{i=1}^n \varepsilon_{2n} + \alpha_2 \sum_{i=1}^n \varepsilon_{3n} \\ \tilde{\alpha}_2 \sum_{i=1}^n \varepsilon_{1n} - \alpha_3 \sum_{i=1}^n \varepsilon_{2n} + \alpha_2 \sum_{i=1}^n \varepsilon_{3n} \end{pmatrix} \\ &= \begin{pmatrix} \tilde{\alpha}_1 \sum_{i=1}^n \varepsilon_{1n} + \tilde{\alpha} \sum_{i=1}^n \tilde{\varepsilon}_n \\ \tilde{\alpha}_2 \sum_{i=1}^n \varepsilon_{1n} + \tilde{\alpha} \sum_{i=1}^n \tilde{\varepsilon}_n \end{pmatrix}, \end{aligned}$$

where $\tilde{\alpha} \sum_{i=1}^n \tilde{\varepsilon}_n = \alpha_2 \sum_{i=1}^n \varepsilon_{3n} - \alpha_3 \sum_{i=1}^n \varepsilon_{2n}$. Hence, the two stochastic trends arise from y_{1n} and a composition of the trends from y_{2n} and y_{3n}. Had we assumed $r = 2$ instead, then there would only be one stochastic trend. From the form of α in this case

$$\alpha = \begin{pmatrix} 0 & 0 \\ \alpha_{21} & \alpha_{22} \\ \alpha_{31} & \alpha_{32} \end{pmatrix},$$

we find that $\alpha_\perp' = (\tilde{\alpha}_1, 0, 0)$ and thus

$$\alpha_\perp' \sum_{i=1}^n \varepsilon_n = \tilde{\alpha}_1 \sum_{i=1} \varepsilon_{1n},$$

implying that for the constructed model of two coupled processes y_{2n} and y_{3n} together with an independent random walk process y_{1n}, the latter provide the stochastic trend of the full $I(1)$ system when $r = 2$.

In the same manner we can construct hypotheses on one-way or bi-directional interaction as well as independence of individual processes in higher dimensional systems and for $r > 1$ by appropriate restrictions on the matrices α and β. Although the restrictions are essentially that $R_\alpha' \alpha = 0$ and/or $R_\beta' \beta = 0$, as seen above, it is often tractable to think of linear restrictions in terms of the matrices H_α, H_β instead. The interpretation of these linear matrices are, that they restrict the span of the columns in α, β to the column space of H_α and H_β respectively,

$$\begin{aligned} \text{sp}(\alpha) &\subset \text{sp}(H_\alpha) \\ \text{sp}(\beta) &\subset \text{sp}(H_\beta). \end{aligned}$$

The discussion regarding stochastic trends generalize to $p > 3$ as well and by forming various hypotheses we can attempt to analyze the origins of any stochastic $I(1)$ trends in the system.

The likelihood ratio tests for the linear restrictions (2.34) for α and β conveniently have $\chi^2_{r(p-s)}$ asymptotical distributions, where $r(p-s)$ denote the degrees of freedom. Here s is the dimension of either ψ or ϕ in (2.34). The likelihood ratio test compares the null against the unrestricted, both under the assumption $\operatorname{rank}(\Pi) = r$, i.e. the test becomes H_0 versus $H(r)$. For restrictions $\beta = H\phi$ (omitting here the H_β notation to simplify), the likelihood is maximized as in (2.26), but the estimated eigenvalues differ. Instead of the (unrestricted β) reduced rank regression $R_{0n} = \alpha\beta' R_{1n} + \hat{\varepsilon}_n$, this becomes $R_{0n} = \alpha\phi' H' R_{1n} + \hat{\varepsilon}_n$ which is in turn solved by the eigenvalue problem

$$|\lambda^* H' S_{11} H - H' S_{10} S_{00}^{-1} S_{01} H| = 0.$$

Denoting the estimated eigenvalues under the restricted reduced rank regression by λ_i^*, the likelihood is maximized by $L_{\max} = |S_{00}| \prod_{i=1}^{r}(1-\lambda_i^*)$ and the likelihood ratio test becomes

$$-2\log Q\big(H_0|H(r)\big) = N\sum_{i=1}^{r}\log\Big(\frac{1-\lambda_i^*}{1-\hat{\lambda}_i}\Big) \sim \chi^2_{r(p-s)}, \tag{2.36}$$

where $\hat{\lambda}_i$ denotes the eigenvalues solving (2.21). The unrestricted β (and α) model have $2pr - r^2$ free parameters whereas under H_0 the model has $pr + rs - r^2$ for the χ^2 distribution (Johansen, 1996). This implies a difference of $pr - rs = r(p-s)$ degrees of freedom. For restrictions on $\alpha = H\psi$ (omitting again the notation H_α to simplify) the end result is similar, but more work is required to obtain the correct form of the eigenvalue problem (2.21). Here we state the result, but refer to Johansen (1996) for the full derivation. For a given matrix $M \in \mathbb{R}^{p\times s}$, define $\bar{M} = M(M'M)^{-1}$ such that $M'\bar{M} = I_s$. Then, with the restriction $\alpha = H\psi$, $H \in \mathbb{R}^{p\times s}$, considering the corrected residuals

$$\begin{aligned}\tilde{R}_{0n} &= R_{0n} - S_{00}H_\perp(H_\perp' S_{00} H_\perp)^{-1} H_\perp' R_{0n}\\ \tilde{R}_{1n} &= R_{1n} - S_{00}H_\perp(H_\perp' S_{00} H_\perp)^{-1} H_\perp' R_{1n}\end{aligned}$$

and defining

$$S_{ij.H_\perp} = N^{-1}\sum_{n=1}^{N}\tilde{R}_{in}\tilde{R}_{jn}, \text{ for } i,j = 0,1,$$

the revised eigenvalue problem becomes

$$|\tilde{\lambda} S_{11.H_\perp} - S_{10.H_\perp}\bar{H}(\bar{H}' S_{00.H_\perp}\bar{H})^{-1}\bar{H}' S_{01.H_\perp}| = 0.$$

Denoting the eigenvalues solving this equation by $\tilde{\lambda}_i$, the likelihood ratio test for H_0 versus $H(r)$ is given, analogously to (2.36), as

$$-2\log Q\big(H_0|H(r)\big) = N\sum_{i=1}^{r}\log\Big(\frac{1-\tilde{\lambda}_i}{1-\hat{\lambda}_i}\Big) \sim \chi^2_{r(p-s)}, \tag{2.37}$$

where again $\hat{\lambda}_i$ denotes the eigenvalues solving (2.21). For a combination of $H_\alpha \in$

$\mathbb{R}^{p\times m}$ and $H_\beta \in \mathbb{R}^{p\times s}$ the eigenvalue problem is revised accordingly, but the degrees of freedom become $r(p-m)+r(p-s)$.

2.3 Cointegration for Continuous Time Diffusions

So far, the theory on cointegration has only dealt with discrete time autoregressive processes. In Kessler and Rahbek (2004) the authors derive a cointegration framework for multivariate continuous time Gaussian diffusion processes. The p-dimensional Ornstein-Uhlenbeck (OU) process

$$dy_t = \Pi y_t dt + DdW_t, t \geq 0 \tag{2.38}$$

for y_0 given, can be considered as a continuous time version of the VAR(1) process, where $\Pi \in \mathbb{R}^{p\times p}$ as before, and $D \in \mathbb{R}^{p\times k}$, $k \leq p$ is such that $\Omega = D'D$ is a positive definite covariance matrix and W_t is a k-dimensional Wiener process.

Kessler and Rahbek (2004) discuss cointegration in the context of the OU process (2.38) with respect to the identifiability of the matrices Π and Ω given discrete time observations in equidistant time intervals, $t_i\delta$, for $i = 0, 1, \ldots, T$, of length δ. The solution to (2.38) is given by

$$y_t = \exp(t\Pi)\Big(y_0 + \int_0^t \exp(-s\Pi)DdW_s\Big), \tag{2.39}$$

where $\exp(\cdot)$ refer to the matrix exponential function for a matrix M

$$\exp(M) = \sum_{i=0}^{\infty} \frac{M^i}{i!}. \tag{2.40}$$

If the matrix Π have full rank p and the real part of the eigenvalues are strictly negative, then the solution (2.39) is stationary and ergodic, but if $\mathrm{rank}(\Pi) = r < p$, then this is not so. However, in this case we can factor $\Pi = \alpha\beta'$ for matrices $\alpha, \beta \in \mathbb{R}^{p\times r}$ as in the discrete time case. Given further assumptions that $|\beta'\alpha| \neq 0$ and the real part of the eigenvalues of $\beta'\alpha$ are negative, we obtain a stationary and ergodic process $\beta' y_t$.

The process (2.39) is only observed at discrete times $0, \delta, 2\delta, \ldots$ and the corresponding discrete time VAR formulation of the solution is

$$y_{t_i} = Ay_{t_{i-1}} + \varepsilon_{t_i}, i = 1, \ldots, T, \tag{2.41}$$

where $\varepsilon_{t_i} \sim \mathcal{N}_p(0, \Sigma)$ is an i.i.d. sequence and

$$A = \exp(\delta\Pi) \tag{2.42}$$

$$\Sigma = f_\Pi(\Omega) = \int_0^\delta \exp(s\Pi)\Omega\exp(s\Pi')ds. \tag{2.43}$$

Rewriting (2.41) in the error correction form (2.8), then

$$\Delta y_{t_i} = (A - I_p)y_{t_{i-1}} + \varepsilon_{t_i} = Py_{t_{i-1}} + \varepsilon_{t_i}, \tag{2.44}$$

for $P = A - I_p = \exp(\delta\Pi) - I_p$. Note that since A is defined by the matrix exponential of $\delta\Pi$, identification of Π depends on the uniqueness of the solution to the matrix exponential function. In Kessler and Rahbek (2004) this is referred to as the *aliasing problem* and an example of the bivariate rotating Brownian motion presents a case of a non-unique solution such that the two solutions Π and $\tilde{\Pi}$ are indistinguishable and thus non-identifiable from discrete observations. However, Kessler and Rahbek (2004) (Lemma 1) state a simple condition on the uniqueness of Π. The eigenvalues of Π are real and no elementary divisor of Π occur more than once, if and only if the $\tilde{\Pi} = \Pi$ is the unique solution to $\exp \delta\tilde{\Pi} = \exp \delta\Pi$ for all real $p \times p$ matrices. Furthermore, a unique solution to the equation $A = \exp(\delta\Pi)$ is ensured if $|A| \neq 0$ and A has no negative eigenvalues.

Kessler and Rahbek (2004) state that if $\Pi \in \mathbb{R}^{p\times p}$ has rank r such that $\Pi = \alpha\beta'$ where $|\beta'\alpha| \neq 0$ and all eigenvalues of $\beta'\alpha$ have negative real parts, then the continuous process (2.38) is cointegrated of order 1 and can be split analogous to the Granger Representation Theorem for discrete time models. In this case, the discrete time version (2.41) of the diffusion has

$$P = \exp(\delta\alpha\beta') - I_p = \alpha\kappa\beta,$$

for the nonsingular matrix $\kappa \in \mathbb{R}^{r\times r}$, composed of α, β and δ. Theorem 1 in Kessler and Rahbek (2004) then states that

$$\operatorname{rank}(P) = \operatorname{rank}(\Pi) = \operatorname{rank}(\alpha\beta) = r,$$

and equally important, for $P = ab'$, with $a, b \in \mathbb{R}^{p\times r}$ of full rank r, then

$$\begin{aligned} \operatorname{sp}(a) &= \operatorname{sp}(\alpha) \\ \operatorname{sp}(b) &= \operatorname{sp}(\beta), \end{aligned}$$

that is, the subspaces for the discrete time matrices a, b coincide with their continuous time counterparts α, β. Although the cointegration matrix Π may not be uniquely identifiable, it is still possible to infer about the rank and the subspaces for α and β in the decomposition of Π and considering the hypotheses in Section 2.2.6, this implies that it is possible to derive meaningful conclusions from discrete observations of the continuous time cointegrated diffusion process (2.38). Given the discrete observations, the procedures for estimation and inference are exactly the discrete time estimation and restrictions outlined above, but conclusions must be transformed between the discrete and continuous time model using the matrix logarithm/exponential function.

MANUSCRIPT I

Oscillating systems with cointegrated phase processes

Jacob Østergaard, Anders Rahbek and Susanne Ditlevsen

Published in Journal of Mathematical Biology.
DOI: 10.1007/s00285-017-1100-2

Abstract

We present cointegration analysis as a method to infer the network structure of a linearly phase coupled oscillating system. By defining a class of oscillating systems with interacting phases, we derive a data generating process where we can specify the coupling structure of a network that resembles biological processes. In particular we study a network of Winfree oscillators, for which we present a statistical analysis of various simulated networks, where we conclude on the coupling structure: the direction of feedback in the phase processes and proportional coupling strength between individual components of the system. We show that we can correctly classify the network structure for such a system by cointegration analysis, for various types of coupling, including uni-/bi-directional and all-to-all coupling. Finally, we analyze a set of EEG recordings and discuss the current applicability of cointegration analysis in the field of neuroscience.

1 Introduction

Since the first scientific discovery of two pendulums synchronizing by Christiaan Huygens in the 17th century, this naturally occurring phenomenon has now been observed in diverse areas such as fireflies synchronizing their flashing behavior, a theatre audience applauding after a show and also in chemical and biological systems, such as the brain and the heart beats of a mother and her fetus, where coupled oscillators appear, see also Pikovsy, Rosenblum, and Kurths (2001). Due to it's pervasive presence, understanding synchronization is of key interest for researchers to understand biological networks, such as the connectivity of the nervous system, circadian rhythms or the cardiovascular system. To a statistician this presents a fascinating challenge of modelling complex behavior in large scale systems and how to

infer the data-generating mechanisms. To this day, synchronization is not fully understood, but has been the centre of research for decades as evident in (Ermentrout, 1985; Kuramoto, 1984; Strogatz, 1987; Strogatz, 2000; Taylor and Holmes, 1998; Winfree, 1967), even the phenomenon of synchronizing pendulums as observed by Huygens, still attracts attention today, see Martens (2013) and Oliveira and Melo (2015). Many innovative ideas have been presented since Winfree (1967) began a mathematical treatment of the subject. When Kuramoto (1984) first presented his model of coupled oscillators, this made a huge impact in the field and spawned a new generation of research on synchronization. Kuramotos model is still considered among one of the most significant advancements in the study of synchronization in oscillating systems as acknowledged by Strogatz (2000), and the study of coupled oscillators still attracts a fair interest from researchers (Ashwin, Coombes, and Nicks, 2016; Burton, Ermentrout, and Urban, 2012; Fernandez and Tsimring, 2014; Ly, 2014; Ly and Ermentrout, 2011).

A long standing problem in neuroscience is to recover the network structure in a coupled system. This could for example be to infer the functional connectivity between units in a network of neurons from multiple extracellularly recorded spike trains, or how traces of EEG signals from different locations on the scalp affect each other, which we will treat in this paper. To the authors knowledge, this challenge is still lacking a sound statistical framework to model and test for interaction in a system, as well as impose statistical hypotheses on the network structure. For this task, cointegration analysis offers a refined statistical toolbox, where detailed information on the connections can be inferred, such as the direction and proportional strength of the coupling. The theory of cointegration was originally conceived by Granger (1981), and has since then also been the subject of intense research, most notably within the field of econometrics. In the monograph by Johansen (1996), the full likelihood theory for linear cointegration models with Gaussian i.i.d. errors is derived, and a framework for estimation and inference on parameters using the quotient test is presented. This well acknowledged framework is popularly termed the Johansen procedure. Even though cointegration analysis has developed from within the field of econometrics, it may potentially be used for different models outside economics, such as biological models in continuous time as we explore here. It has also been applied in climate analysis, see Schmidt, Johansen, and Thejll (2012).

In this paper, we demonstrate how to apply cointegration analysis to a system of linearly phase coupled oscillating processes. To display the applicability of the method, we present a simulation experiment, where we present a statistical analysis of phase coupled systems with varying network structures, including uni-/ bi-directional and all-to-all couplings. We show that we can identify the proportional coupling strengths and directions given by the estimated *cointegration matrix* parameter. Our work is inspired by Dahlhaus and Neddermeyer (2012), which also introduces cointegration analysis as a statistical toolbox to neuroscientists and new challenges for researchers in cointegration theory. However, in contrast to Dahlhaus and Neddermeyer (2012), we incorporate the fact that we are dealing with continuous systems and also ensure that the cointegration property of the system is well posed as a linear structure. This approach assures that the conclusion on the interaction in the data is accurate in terms of cointegration.

The paper is composed as follows. In Section 2 we define a class of phase coupled oscillators, in Section 3 we highlight some cointegration theory for the analysis including an extension to discretely observed, continuous time models. In Section 4

we present a statistical analysis of linearly phase coupled oscillating systems and in Section 5 we analyze EEG recordings from an epileptic subject experiencing a seizure, previously analyzed by Shoeb (2009). We discuss the model and findings, conclude on the research and give an outlook of the future direction of the research in Section 6. Technical details are presented in the appendix.

Throughout we use the following notation and conventions: unless explicitly stated otherwise, time $t \in [0, \infty)$ is assumed continuous, and the process $(x_t, y_t)'$ is assumed observed with corresponding polar coordinates $(\phi_t, \gamma_t)'$. Here $'$ denotes transposition. For a $p \times r$ matrix M, with $r \leq p$, we denote the orthogonal complement $M_\perp$, a $p \times (p-r)$ matrix such that $M_\perp' M = 0$ (zero matrix). Also denote by $\text{sp}(A)$ the subspace spanned by the columns of a matrix A, and let $\text{rank}(A)$ denote the rank of the matrix, i.e., the dimension of $\text{sp}(A)$.

2 Oscillating systems

Studying biological rythms corresponds to studying systems of periodical processes. Intuitively we define a single *oscillator* as a continuous time bi-variate process $z_t = (x_t, y_t)' \in \mathbb{R}^2$, $t \in [0, \infty)$, such that z_t revolve around some arbitrary center. Such a process can be derived from an equivalent process in polar coordiantes $(\phi_t, \gamma_t)'$, where $\phi_t \in \mathbb{R}$ is the *phase* process and $\gamma_t \in \mathbb{R}$ is the *amplitude* process, such that

$$\begin{aligned} x_t &= \gamma_t \cos(\phi_t) \\ y_t &= \gamma_t \sin(\phi_t). \end{aligned} \tag{I.1}$$

We then define the process z_t to be an oscillator if the phase process has a monotonic trend.

2.1 Defining a class of coupled oscillators

Definition (I.1) naturally extends to a system of *coupled stochastic oscillators*, where we observe p oscillators that interact, i.e., $z_t \in \mathbb{R}^{2p}$. Define a class of oscillators with phase ($\phi_t \in \mathbb{R}^p$) and amplitude ($\gamma_t \in \mathbb{R}^p$) processes given by the multivariate stochastic differential equations (SDE)

$$d\phi_t = f(\phi_t, \gamma_t)dt + \Sigma_\phi dW_t^\phi \tag{I.2}$$

$$d\gamma_t = g(\phi_t, \gamma_t)dt + \Sigma_\gamma dW_t^\gamma, \tag{I.3}$$

where $f, g : \mathbb{R}^{2p} \to \mathbb{R}^p$ are real valued vector functions, possibly depending on both ϕ_t, γ_t or constant, dW_t^ϕ, dW_t^γ are multivariate standard Wiener processes and $\Sigma_\phi, \Sigma_\gamma \in \mathbb{R}^{p \times p}$ such that $\Sigma_i \Sigma_i'$ is a positive semi-definite covariance matrix for $i = \phi, \gamma$. Assume the properties of (I.2) and (I.3) are such that

$$\gamma_t \in \mathbb{R}_+^p \text{ for } t \in [0, \infty) \tag{I.4}$$

and

$$\mathbb{E}[\phi_{kt}] \text{ is monotonically increasing as a function of } t \text{ for each } k = 1, \ldots, p, \tag{I.5}$$

where $\mathbb{E}[\cdot]$ denotes the mean. Since $\gamma_t = (\gamma_{1t}, \ldots, \gamma_{pt})'$ are interpreted as the amplitudes of the individual oscillators, eq. (I.4) is a natural assumption and eq. (I.5) ensures that the individual oscillators actually revolve (anti-clockwise) around the center and that they are not "stuck" in some part of the phase space, i.e., their *angular velocities* are positive. Note that we have defined the phase-trend as positive, corresponding to counter-clockwise rotation in accordance with the standard interpretation of the phase. However, for a negative trending process, one can either look at $-\phi_t$ or simply interpret rotations as clockwise.

To emphasize the implication of inducing interaction in a system, for the data generating process (DGP) in the xy-plane, we derive a DGP from (I.2)-(I.3), see appendix A. Assuming that $\Sigma_\phi = \operatorname{diag}(\sigma_1^\phi, \ldots, \sigma_p^\phi)$ and $\Sigma_\gamma = \operatorname{diag}(\sigma_1^\gamma, \ldots, \sigma_p^\gamma)$ we find that

$$\begin{aligned} d\begin{pmatrix} x_{kt} \\ y_{kt} \end{pmatrix} = &\begin{pmatrix} -\frac{1}{2}(\sigma_k^\phi)^2 & -f_k(\phi_t, \gamma_t) \\ f_k(\phi_t, \gamma_t) & -\frac{1}{2}(\sigma_k^\phi)^2 \end{pmatrix} \begin{pmatrix} x_{kt} \\ y_{kt} \end{pmatrix} dt + \begin{pmatrix} 0 & -\sigma_k^\phi \\ \sigma_k^\phi & 0 \end{pmatrix} \begin{pmatrix} x_{kt} \\ y_{kt} \end{pmatrix} dW_{kt}^\phi \\ &+ \frac{g_k(\phi_t, \gamma_t) + \sigma_k^\gamma \sigma_k^\phi}{\sqrt{x_{kt}^2 + y_{kt}^2}} \begin{pmatrix} x_{kt} \\ y_{kt} \end{pmatrix} dt + \frac{\sigma_k^\gamma}{\sqrt{x_{kt}^2 + y_{kt}^2}} \begin{pmatrix} x_{kt} \\ y_{kt} \end{pmatrix} dW_{kt}^\gamma. \end{aligned} \tag{I.6}$$

Hence, with the definitions (I.2)-(I.5) we have introduced a general class of coupled oscillators, where the specifications of f and g define the properties of the system, such as interaction. This broad definition of oscillating systems covers among others the Kuramoto model, see Example 2.5 below and other standard oscillators such as the FitzHugh-Nagumo and the Duffing oscillator. In this paper we will analyze *phase coupled* oscillators, and therefore we assume that $g_k(\phi_t, \gamma_t) = g_k(\gamma_{kt})$, such that there is no feedback from the phase process ϕ_t into the amplitude and the k'th amplitude is not dependent on the rest. Hence, interaction in the system is solely through $f(\phi_t, \gamma_t)$, such that the phase processes are attracted by some interdependent relation.

2.2 Linear coupling

The arbitrary function f enables us to choose any transformation of the variables to obtain a coupled system, including unidirectional coupling between phases or periodic forcing of the system if we extend f to depend on t as well, intermittent synchronization dependent on a threshold in process differences, etc.

Studying the general case where $f(\phi_t, \gamma_t)$ is nonlinear in ϕ_t and γ_t is a complex exercise. In this paper we restrict ourselves to models where f is composed of a linear mapping of ϕ_t and a function of γ_t, with components,

$$f_k(\phi_t, \gamma_t) = \sum_{j=1}^{p} \Pi_{kj}(\phi_{jt} - \omega_j) + h(\gamma_{kt}), \quad \text{for } k = 1, \ldots, p \tag{I.7}$$

for a real matrix $\Pi \in \mathbb{R}^{p \times p}$ and constant vector $\omega = (\omega_1, \ldots, \omega_p)' \in \mathbb{R}^p$. With this restriction, the interaction between oscillators is linear in the phase, and the k'th oscillator is only dependent on the intrinsic amplitude γ_{kt} through $h(\gamma_{kt})$. We will refer to such a system as *linearly phase coupled*.

Although we impose the linear restriction Π on the interaction between phases, we can still model a broad set of coupling structures as we show with examples below. Since the interaction is given by $\Pi\phi_t$, we note that the *coupling strength* in the system is given as the absolute values of the entries of Π and that row k of Π define how oscillator k depends on the rest. Note also that ω defines the attracting state for the phase relations, see Example 2.3 below. Normally $h(\gamma_{kt})$ is restricted to a constant, but in Section 4 we will relax this and investigate systems where $h(\gamma_{kt})$ is only approximately linear and has a sufficiently low variance. This implies a misspecified model, but as we will show, we can still identify the coupling structure, although inference on $h(\gamma_{kt})$ itself is less meaningful.

2.3 Example: Linearly phase coupled system with a degenerate γ_t process

Let f be defined as in (I.7) and assume that γ_t is a constant (positive) process such that $h(\gamma_{kt}) = \mu_k > 0$. Then f is of the form

$$f(\phi_t) = \Pi(\phi_t - \omega) + \mu, \tag{I.8}$$

where $\omega, \mu \in \mathbb{R}^p$ are constant vectors. For reduced rank matrices Π (I.2) is a continuous time cointegrated process (see Section 3) and f admits a linearly phase coupled system with intrinsic rotating frequencies μ. Note that if $\Pi = 0$ then there is no interaction in the system, and the individual oscillators will rotate according to their own $\mu_k > 0$, and we refer to the system as *independent*.

The linear specification $\Pi(\phi_t - \omega)$ implies that at most one attracting point can exist. As an illustration of this, assume a system composed of two coupled oscillators, with

$$\Pi(\phi_t - \omega) = \begin{pmatrix} -\alpha_1 & \alpha_1 \\ \alpha_2 & -\alpha_2 \end{pmatrix} \begin{pmatrix} \phi_{1t} - \omega_1 \\ \phi_{2t} - \omega_2 \end{pmatrix} = \begin{pmatrix} -\alpha_1 \\ \alpha_2 \end{pmatrix} \Big((\phi_{1t} - \phi_{2t}) - (\omega_1 - \omega_2) \Big).$$

where $0 < \alpha_1 + \alpha_2 < 2$. Since $\omega^* = \omega_1 - \omega_2$ define an attracting state of the phase difference $\phi_{1t} - \phi_{2t}$, then with $\omega^* = 0$ the system is attracted towards being *in-phase*, whereas $\omega^* = \pi$ would imply that the system is attracted towards being in *anti-phase*. Considering that neither α_1, α_2 or ω^* depend on time, the system cannot switch to a different attracting regime.

To illustrate possible coupling structures, consider again the system of two oscillators and assume that $\omega = 0$. Then with $\alpha_2 = 0$ and $\alpha_1 \neq 0$ the coupling between ϕ_{1t}, ϕ_{2t} is *uni-directional* $\phi_{2t} \to \phi_{1t}$ where the arrow $\to$ denote the direction of interaction. Likewise, if $\alpha_1 = 0$ and $\alpha_2 \neq 0$ then $\phi_{1t} \to \phi_{2t}$. However, if both $\alpha_1, \alpha_2 \neq 0$ then $\phi_{2t} \leftrightarrow \phi_{1t}$ and the coupling is *bi-directional*. In general, if ϕ_{kt} appears in the expression $f_l(\phi_t)$ for oscillator $l \neq k$, then $\phi_{kt} \to \phi_{lt}$. If the opposite is true, then $\phi_{lt} \to \phi_{kt}$ and if both directions exist, then $\phi_{lt} \leftrightarrow \phi_{kt}$. For $f_k(\phi_t) = 0$ oscillator k is (oneway) independent from the rest, but it can still possibly influence others.

For systems where γ_t is a degenerate process, then $\Sigma_\gamma = 0$ and $g(\phi_t, \gamma_t) = 0$. With $\sigma_k^\phi = \sigma_k$ then (I.6) simplifies to

$$d \begin{pmatrix} x_{kt} \\ y_{kt} \end{pmatrix} = \begin{pmatrix} -\frac{1}{2}\sigma_k^2 & -f_k(\phi_t) \\ f_k(\phi_t) & -\frac{1}{2}\sigma_k^2 \end{pmatrix} \begin{pmatrix} x_{kt} \\ y_{kt} \end{pmatrix} dt + \begin{pmatrix} 0 & -\sigma_k \\ \sigma_k & 0 \end{pmatrix} \begin{pmatrix} x_{kt} \\ y_{kt} \end{pmatrix} dW_k, \tag{I.9}$$

where $f_k(\phi_t) = \sum_j \Pi_{kj}\phi_{jt} + \mu_k$. Note that if $\Pi = 0$ then (I.8) is simply a constant trend and hence (I.9) is a rotating process. One can show that the eigenvalues of the deterministic drift matrix in (I.9) in this case are complex conjugates, $-\frac{\sigma^2}{2} \pm i\mu$, where $i = \sqrt{-1}$, implying that the solutions to (I.9) oscillate for $\mu \neq 0$. The oscillations are damped by the negative real part, but sustained by the noise term.

When γ_t is a constant vector process the properties of the system are fully identified by (I.2). Furthermore, if the noise level of the phases Σ_ϕ is sufficiently small, we can use the Hilbert transform[1] to derive the phase process ϕ_t from observations of either x_t or y_t. This is a commonly used technique in signal processing and has been applied to oscillating systems as well, see Dahlhaus and Neddermeyer (2012) and Pikovsy, Rosenblum, and Kurths (2001). For systems where ϕ_t is very noisy, this method is less applicable.

2.4 Example: Winfree oscillator

Let $g_k(\phi, \gamma) = (\kappa_k - \gamma_k)\gamma_k^2$ for a vector $\kappa \in \mathbb{R}_+^p$ and $f_k(\phi, \gamma) = \sum_{j=1}^p \Pi_{kj}\phi_j + \gamma_k$ for $\Pi \in \mathbb{R}^{p\times p}$ such that

$$
\begin{aligned}
d\gamma_{kt} &= (\kappa_k - \gamma_{kt})\gamma_{kt}^2 dt + \sigma_k^\gamma dW_{kt}^\gamma \\
d\phi_{kt} &= \Big(\sum_{j=1}^p \Pi_{kj}\phi_j + \gamma_{kt}\Big)dt + \sigma_k^\phi dW_{kt}^\phi.
\end{aligned}
$$

With these definitions (I.6) becomes

$$
\begin{aligned}
d\begin{pmatrix} x_{kt} \\ y_{kt} \end{pmatrix} = &\begin{pmatrix} (\kappa_k - \gamma_{kt})\gamma_{kt} + \gamma_{kt}^{-1}\sigma_k^\gamma\sigma_k^\phi - \frac{1}{2}(\sigma_k^\phi)^2 & -\Big(\sum_{j=1}^p \Pi_{kj}\phi_j + \gamma_{kt}\Big) \\ \Big(\sum_{j=1}^p \Pi_{kj}\phi_j + \gamma_{kt}\Big) & (\kappa_k - \gamma_{kt})\gamma_{kt} + \gamma_{kt}^{-1}\sigma_k^\gamma\sigma_k^\phi - \frac{1}{2}(\sigma_k^\phi)^2 \end{pmatrix}\begin{pmatrix} x_{kt} \\ y_{kt} \end{pmatrix} dt \\
&+ \begin{pmatrix} 0 & -\sigma_k^\phi \\ \sigma_k^\phi & 0 \end{pmatrix}\begin{pmatrix} x_{kt} \\ y_{kt} \end{pmatrix} dW_{kt}^\phi + \begin{pmatrix} \gamma_{kt}^{-1}\sigma_k^\gamma & 0 \\ 0 & \gamma_{kt}^{-1}\sigma_k^\gamma \end{pmatrix}\begin{pmatrix} x_{kt} \\ y_{kt} \end{pmatrix} dW_{kt}^\gamma.
\end{aligned}
\tag{I.10}
$$

This example is taken from Winfree (2001) and extended with noise and phase interaction, and therefore we will refer to (I.10) as the (noisy) *Winfree oscillator*. Note that the formulation of $d\gamma_{kt}$ implies that the amplitude fluctuates around κ_k. Due to this, we can for sufficiently small noise Σ_γ insist that $\gamma_{kt} \approx \kappa_k$ for $k = 1, \ldots, p$ and therefore analyze the Winfree oscillator using the cointegration toolbox, assuming a constant γ_t in $d\phi_t$. In Section 4 we analyze the range of noise, Σ_γ, where the cointegration analysis still performs well.

2.5 Example: Kuramoto model

Choose $f(\phi_t, \gamma_t)$ such that

$$
f_k(\phi_t, \gamma_t) = f_k(\phi_t) = \frac{1}{p}\sum_{j=1}^p K_{kj}\sin(\phi_{jt} - \phi_{kt}) + \mu_i, \quad k = 1, \ldots, p, \tag{I.11}
$$

[1] The Hilbert transform of a signal x_t is defined as $H(x_t) = \pi^{-1}\text{p.v.}\int_{-\infty}^{\infty}\frac{x_\tau}{t-\tau}d\tau = -\pi^{-1}\lim_{\varepsilon\to 0}\int_\varepsilon^\infty \frac{x_{t+\tau}-x_{t-\tau}}{\tau}d\tau$, where p.v. $\int_{-\infty}^{\infty}$ denotes the principal value integral.

then (I.2) is the Kuramoto model extended with a stochastic noise term, for phase coupled oscillators, where K_{kj} denotes the coupling strength between the k'th and j'th oscillators. In the classic version, $K_{kj} = K \; \forall k, j$, such that for a certain threshold K_c, then with $K > K_c$ the oscillators exhibit synchronization. For an arbitrary γ_t process we cannot simplify (I.6), but with a degenerate γ_t we obtain the same expression as in (I.9) with $f_k(\phi_t)$ as in (I.11).

For the Kuramoto model f is a nonlinear function, hence it is not directly applicable to a standard cointegration analysis where f is assumed linear. To emphasize this fact, consider the special case $p = 2$, where the Kuramoto model is particularly simple and (I.11) can be written explicitly as,

$$\begin{aligned} f(\phi_t) &= \frac{1}{2}\begin{pmatrix}\alpha_1 \sin(\phi_{2t} - \phi_{1t}) \\ \alpha_2 \sin(\phi_{1t} - \phi_{2t})\end{pmatrix} + \mu = \frac{1}{2}\begin{pmatrix}-\alpha_1 \\ \alpha_2\end{pmatrix} \sin(\phi_{1t} - \phi_{2t}) + \mu \\ &= \frac{1}{2}\begin{pmatrix}-\alpha_1 \\ \alpha_2\end{pmatrix} \sin(\beta'\phi_t) + \mu. \end{aligned}$$

where $\beta' = (1, -1)$ and $(\alpha_1, \alpha_2) = (K_{12}, K_{21})$. If $\phi_{1t} \approx \phi_{2t}$ at $t = 0$ and the values of α_1, α_2 are large enough, then $\phi_{1t} \approx \phi_{2t} \; \forall t$, such that $\beta'\phi_t \approx 0$ and we can write a crude linear approximation of the sine function: $\sin(\beta'\phi_t) \approx \beta'\phi_t$, such that

$$f(\phi_t) \approx \frac{1}{2}\begin{pmatrix}-\alpha_1 \\ \alpha_2\end{pmatrix} \beta'\phi_t + \mu = \frac{1}{2}\begin{pmatrix}-\alpha_1 & \alpha_1 \\ \alpha_2 & -\alpha_2\end{pmatrix} \phi_t + \mu. \tag{I.12}$$

This is a coarse, but linear, approximation of the Kuramoto model and we can perform a cointegration analysis assuming this approximation is satisfactory. However, one must be cautious with this approximation. Consider $\sin(\beta'\phi_t)$, when $\beta'\phi_t = \phi_{1t} - \phi_{2t} \approx \pi$. In this case $\sin(\beta'\phi_t) \approx \pi - \beta'\phi_t$, and hence

$$f(\phi_t) \approx \frac{1}{2}\begin{pmatrix}-\alpha_1 \\ \alpha_2\end{pmatrix} (\pi - \beta'\phi_t) + \mu = \frac{1}{2}\begin{pmatrix}\alpha_1 & -\alpha_1 \\ -\alpha_2 & \alpha_2\end{pmatrix} \phi_t + \mu + \begin{pmatrix}-\alpha_1 \\ \alpha_2\end{pmatrix} \pi, \tag{I.13}$$

and we see that not only do we add a term with π, but the interaction also reverses sign. Recall that $0 < \alpha_1 + \alpha_2 < 2$ which implies a stationary relation in the system in (I.12), see Section 3.2. In (I.13) this condition is reversed, in the sense that $-2 < \alpha_1 + \alpha_2 < 0$ will imply stationarity. If $0 < \alpha_1 + \alpha_2 < 2$, (I.13) leads to an explosive system, which is not covered in this paper. Therefore, an essential requirement for an approximation of the Kuramoto model is a regime switching ability of (I.2). For a model with this property, we propose that cointegration analysis on a piecewise linear approximation of the Kuramoto model does make sense and can lead to correct conclusions regarding the network structure. In this paper we will not deal with non-linear cointegration of oscillating systems, but leave this direction open for future research. For a statistical analysis of nonlinear cointegrated systems of the form $\alpha_t\beta'$, i.e. time varying, or regime switching α coefficients, see Bec and Rahbek (2004) and Kristensen and Rahbek (2013).

Note that with a general coupling constant $K_{kj} = K$, then the simple linear approximation to the Kuramoto model around $\phi_{jt} - \phi_{kt} \approx 0$ is

$$\frac{K}{p}\sum_{j=1}^{p}\begin{pmatrix}\sin(\phi_{jt} - \phi_{1t}) \\ \vdots \\ \sin(\phi_{jt} - \phi_{pt})\end{pmatrix} \approx \frac{K}{p}\begin{pmatrix}-(p-1) & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & -(p-1)\end{pmatrix} \phi_t. \tag{I.14}$$

3 Cointegration

Cointegration theory was originally developed for discrete time processes, however the ubiquitous use of continuous time models has inspired development of continuous time cointegration theory, see Kessler and Rahbek (2004) and Kessler and Rahbek (2001). In order to present cointegration analysis as a framework for phase-processes, we therefore review some background on *discrete* time processes before entering into continuous time cointegrated models. The first part of this section is based on Johansen (1996) and Lütkepohl (2005).

3.1 Integrated process

Assume that ϕ_n is a discrete time vector autoregressive process,

$$\phi_n = A\phi_{n-1} + \mu + \varepsilon_n, \tag{I.15}$$

where $A \in \mathbb{R}^{p\times p}$, ε_n is a Gaussian white noise and $\mu \in \mathbb{R}^p$ is a deterministic term. The *characteristic polynomial* for (I.15) is the determinant of $I_p - A\zeta$ for $\zeta \in \mathbb{C}$, where I_p is the p-dimensional identity matrix. If the roots of the characteristic polynomial are all outside the unit circle, then the initial values of ϕ_n can be given a distribution such that ϕ_n is stationary, see Johansen (1996).

If the characteristic polynomial of (I.15) contains one or more roots at $\zeta = 1$, then there is no stationary solution of ϕ_n, and we say that the process is *integrated*. In particular, see Johansen (1996), $P = A - I_p$ will have reduced rank $r < p$ and can be written as $P = ab'$ with a, b $(p \times r)$ matrices of rank r. Moreover, the process ϕ_n is integrated of order one, $I(1)$ with r cointegrating relations $b'\phi_n$ under regularity conditions presented in Section 3.2. Note that the order of integration is a stochastic property and hence including deterministic terms in a model does not change the order of integration.

In this paper we will only deal with $I(1)$ processes, so when we refer to ϕ_n as integrated, we implicitly mean that ϕ_n is integrated of order 1.

3.2 Cointegrated process

Let $\phi_n = (\phi_{1n}, ..., \phi_{pn})' \in R^p$ and rewrite (I.15) with $P = A - I_p$ as

$$\Delta\phi_n = P\phi_{n-1} + \mu + \varepsilon_n. \tag{I.16}$$

As already noted if $\det(I - A\zeta) = 0$ implies $|\zeta| > 1$ then ϕ_n has a stationary representation (as an $I(0)$ process). In particular, P has full rank p and all linear combinations of ϕ_n are stationary. If the $(p \times p)$-dimensional matrix P has reduced rank $r < p$ then $P = ab'$ with a, b, $p \times r$ dimensional matrices of rank r. Moreover, the process ϕ_n is integrated of order one, $I(1)$ with r cointegrating stationary relations $b'\phi_n$ provided $\rho(I_r + b'a) < 1$ with $\rho(\cdot)$ denoting the spectral radius. This we refer to as the $I(1)$ conditions in the following.

Note that if $r = 0$ the process ϕ_n is $I(1)$ with no cointegration, while if $r = p$ (and $\rho(A) < 1$) then ϕ_n is $I(0)$, or p stationary linear combinations exist. Under the reduced rank r, the system is written as,

$$\Delta\phi_n = ab'\phi_{n-1} + \mu + \varepsilon,$$

with b containing the r cointegration vectors and a the *loadings* or *adjustment coefficients*. Note that the entries of a and b are not uniquely identified, since we can use any non-singular transformation to obtain similar results. Rather we identify the subspaces $\text{sp}(a), \text{sp}(b) \in \mathbb{R}^r$, that is, the subspaces spanned by the columns of a, b, where we use the normalization

$$b^* = b(c'b)^{-1}, \quad \text{with } c = (I_r, 0_{p-r\times r})'$$

of b in order to identify parameters uniquely. Furthermore, let $m_\perp$ denote the matrix such that $\text{sp}(m_\perp)$ is orthogonal to $\text{sp}(m)$, then a necessary condition for an $I(1)$ process is that $|a'_\perp b_\perp| \neq 0$. For more on estimation and inference in cointegration models, see Appendix B.

3.3 Continuous time cointegrated models

Kessler and Rahbek (2004) and Kessler and Rahbek (2001) derive a cointegration theory for continuous time models, and conclude that for a discretely observed process, using conventional methods for discrete time generally apply to inference on continuous time parameters. Consider (I.2) with f as in (I.8) and for simplicity $\omega = 0$. This is a p-dimensional Ornstein-Uhlenbeck process. The exact solution is

$$\phi_t = \exp(t\Pi)\Big[\phi_0 + \int_0^t \exp(-s\Pi)\mu ds + \int_0^t \exp(-s\Pi)\Sigma dW_s\Big]. \tag{I.17}$$

Note that for the solution (I.17) to be stationary, then Π must be a full rank matrix, and all eigenvalues must have a strictly negative real part. This implies that if Π is *not* of full rank, then ϕ_t is necessarily not stationary.

Assuming discrete observations of (I.17) at equidistant timepoints $t_1 = 0 < t_2 < \cdots < t_N = T$ with timestep $\delta = t_n - t_{n-1}$, the corresponding vector autoregressive process is

$$\phi_{t_n} = \exp(\delta\Pi)\phi_{t_{n-1}} + \delta\mu + \varepsilon_{t_n}, \tag{I.18}$$

such that the difference process can be written as

$$\Delta\phi_{t_n} = \phi_{t_n} - \phi_{t_{n-1}} = \delta P\phi_{t_{n-1}} + \delta\mu + \varepsilon_{t_n},$$

where $\varepsilon \sim \mathcal{N}(0, \Omega)$ and

$$\begin{aligned} P &= \delta^{-1}(\exp(\delta\Pi) - I_p) \\ \Omega &= \int_0^\delta \exp(s\Pi)\Sigma\Sigma'\exp(s\Pi')ds. \end{aligned} \tag{I.19}$$

Results (I.18) and (I.19) hold in general for multivariate processes. Thus, to obtain an estimate for the continuous time matrix, $\hat{\Pi}$, from the discrete time estimate $\hat{P}$, a

logarithmic transformation involving $\hat{P}$ is required

$$\hat{\Pi} = \delta^{-1} \log\left(\delta \hat{P} + I_p\right). \tag{I.20}$$

For a univariate process (I.20) is unique, however this is not the case for a multivariate process, due to the non-uniqueness of the multivariate logarithm. Because of this, we cannot uniquely identify $\hat{\Pi}$, even though we have a unique estimate $\hat{P}$.

For a continuous time process ϕ_t, however, Kessler and Rahbek (2004) and Kessler and Rahbek (2001) conclude that this is cointegrated if and only if the discretely observed process (I.18) is cointegrated. In this case P is of reduced rank, and can be decomposed $P = ab'$ with $a, b \in \mathbb{R}^{p\times r}$ of full rank $r \leq p$. However, it also holds that

$$P = ab' = \alpha\xi\beta' \tag{I.21}$$

for a non-singular matrix $\xi = (\beta'\alpha)^{-1}\Big(\exp(\delta\beta'\alpha) - I_r\Big) \in \mathbb{R}^{r\times r}$ and matrices $\alpha, \beta \in \mathbb{R}^{p\times r}$, such that given weak conditions on the sampling timestep δ (see below), the following relations hold

$$\begin{aligned} \operatorname{rank}(P) &= \operatorname{rank}(\Pi) = r \\ \operatorname{sp}(\alpha) &= \operatorname{sp}(a) \\ \operatorname{sp}(\beta) &= \operatorname{sp}(b), \end{aligned}$$

see Kessler and Rahbek (2001) and Kessler and Rahbek (2004). Hence, for continuous time cointegrated processes, we can infer on the number of cointegration relations ($\operatorname{rank}(\Pi) = r$) from discrete time observations, and also identify the subspaces spanned by the columns of α and β. Note however that due to the unidentified scaling ξ, we can only identify the subspaces, but not the parameters α, β themselves. They are only unique up to a scaling (ξ), even though we have imposed the normalization (I.23) and thus uniquely identified a and b.

In the numerical part, we will refer to estimates of α and β, implicitly referring to the *discrete time* estimates. In terms of subspaces, there is no difference between the discrete and continuous time, but in order to interpret the *continuous time* Π matrix, one must translate the discrete estimate to a continuous estimate using (I.19).

It is important to note that when working with continuous time models, one must be careful with regard to the relation (I.19) between discrete and continuous time and the sampling timestep δ. Kessler and Rahbek (2004) refer to this issue as the *embedding problem*, and to ensure that the continuous time model is appropriate, one must check for $\exp(\delta\Pi)$ in (I.18) that it is non-singular, i.e., $|\exp(\delta\Pi)| \neq 0$, and that it has no negative eigenvalues. If this is the case and the underlying process is in fact cointegrated, the results above hold.

3.4 Likelihood ratio test for $\operatorname{rank}(\Pi) = r$

Consider discrete observations $(\phi_{t_1}, \ldots, \phi_{t_N})$ from the continuous process (I.17) and denote by H_r the hypothesis $H_r : \operatorname{rank}(\Pi) \leq r$ for $r = 0, \ldots, p$. Then the set of hypotheses $H_0, \ldots, H_r$ is nested,

$$H_0 \subseteq H_1 \subseteq \cdots \subseteq H_p,$$

and H_p correspond to the unrestricted model. The likelihood ratio test (LRT) that compare H_r and H_p is applied sequentially for $r = 0, 1, \ldots, p-1$ and continued until H_r against H_p cannot be rejected, and thus determine the number of cointegrating relations for ϕ_t. The LRT statistic is given by

$$-2 \log Q(H_r|H_p) = \sum_{i=r+1}^{p} \hat{\lambda}_i, \tag{I.22}$$

where $\hat{\lambda}_i$ are the solutions to the eigenvalue problem (I.49), see Appendix B. The asymptotic distribution of (I.22) is non-standard and therefore it must be simulated. Here, to also improve on small-sample performance we perform *bootstrap* simulations as presented by Cavaliere, Rahbek, and Taylor (2012), in order to determine critical values. Specifically, given the data $\{\phi_{t_n}\}_{n=1}^{N}$ bootstrap sequences $\{\phi_{t_n}^{*(m)}\}_{n=1}^{N}$ for $m = 1, \ldots, M$ are simulated and for each sequence the LRT statistic $\text{LRT}^{*(m)}$ is re-computed. The empirical quantiles of $\{\text{LRT}^{*(m)}\}_{m=1}^{M}$ are then used for testing. With r determined, $\hat{\beta}$ is given by the r eigenvectors corresponding to $\hat{\lambda}_i, i = 1, \ldots, r$ and the parameter estimates $\hat{\alpha}, \hat{\mu}, \hat{\Sigma}$ follow by ordinary least squares estimation as outlined in Appendix B.

3.5 Inference for α and β

Since we identify subspaces for α and β, then a normalization is necessary to identify the parameters uniquely. If $\hat{\beta}$ is known, then $\hat{\alpha}$ follows by OLS. Hence, if we impose a normalization on $\hat{\beta}$, we can identify all parameters. A common normalization, see Johansen (1996), is

$$\hat{\beta} = \tilde{\beta}(c'\tilde{\beta})^{-1},$$

where $c = (I_r, 0_{p-r \times r})'$ is a $p \times r$ matrix and $\tilde{\beta}$ is any version of the r eigenvectors corresponding to the r largest eigenvalues. This ensures that

$$\hat{\beta} = \begin{pmatrix} I_r \\ \tilde{\beta}_{p-r,r} \end{pmatrix}. \tag{I.23}$$

Extending the idea of normalization to restrictions for α, β, we can impose such under the hypothesis H_r. Assume that $\text{rank}(\Pi) = r$ and that the parameters $\alpha \in \mathbb{R}^{p \times r}$, $\beta \in \mathbb{R}^{p \times r}$, $\mu \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}^{p \times p}$ are all unrestricted within their corresponding subspaces, except for normalization (I.23). Possible hypotheses for α, β are linear restrictions as given by

$$H_\alpha : \alpha = A\psi$$
$$H_\beta : \beta = B\xi,$$

where $A \in \mathbb{R}^{p \times m}$, $\psi \in \mathbb{R}^{m \times r}$, $B \in \mathbb{R}^{p \times s}$, $\xi \in \mathbb{R}^{s \times r}$. The known matrices A and B represent the linear hypotheses and ψ and ξ are parameters to be estimated. It is also possible to combine the hypotheses for α and β and we denote this $H_{\alpha,\beta}$.

As an example, assume a system of 3 oscillators $\phi_t = (\phi_{1t}, \phi_{2t}, \phi_{3t})'$ with $r = 1$. If we believe that ϕ_{3t} is independent of ϕ_{1t} and ϕ_{2t}, we can specify the hypothesis

$$H_\alpha : A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}, \tag{I.24}$$

such that

$$\Pi_A = \alpha_A \beta' = \begin{pmatrix} \psi_1 \\ \psi_2 \\ 0 \end{pmatrix} (\beta_1, \beta_2, \beta_3) = \begin{pmatrix} \psi_1\beta_1 & \psi_1\beta_2 & \psi_1\beta_3 \\ \psi_2\beta_1 & \psi_2\beta_2 & \psi_2\beta_3 \\ 0 & 0 & 0 \end{pmatrix}.$$

This restriction imply that ϕ_{1t} and ϕ_{2t} do not contribute to the dynamics of ϕ_{3t}, and hence that the latter is independent.

If we want to investigate a possible 1:1 coupling between ϕ_{1t} and ϕ_{2t}, we can specify

$$H_\beta : B = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \tag{I.25}$$

and obtain

$$\Pi_B = \alpha \beta'_B = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} (\eta, -\eta, 0) = \begin{pmatrix} \alpha_1\eta & -\alpha_1\eta & 0 \\ \alpha_2\eta & -\alpha_2\eta & 0 \\ \alpha_3\eta & -\alpha_3\eta & 0 \end{pmatrix}.$$

Note however, that under H_β the interaction between ϕ_{1t} and ϕ_{2t} also influence ϕ_{3t} if $\alpha_3 \neq 0$. Hence, the system admits the relations $\phi_{1t} \leftrightarrow \phi_{2t}$, $\phi_{1t} \rightarrow \phi_{3t}$ and $\phi_{2t} \rightarrow \phi_{3t}$, where the restriction $\beta_3 = 0$ implies that the last two relations are unidirectional.

If we believe that ϕ_{1t} and ϕ_{2t} are bi-directionally coupled, $\phi_{1t} \leftrightarrow \phi_{2t}$, but ϕ_{3t} is independent and does not contribute to either ϕ_{1t} nor ϕ_{2t}, we can phrase this hypothesis as a combination of (I.24) and (I.25). This leads to the restricted matrix

$$\Pi_{A,B} = \alpha_A \beta'_B = \begin{pmatrix} \psi_1 \\ \psi_2 \\ 0 \end{pmatrix} (\eta, -\eta, 0) = \begin{pmatrix} \psi_1\eta & -\psi_1\eta & 0 \\ \psi_2\eta & -\psi_2\eta & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Other hypotheses, such as equal or proportional coupling strength or $l : n$ coupling, can be specified by appropriately designing the matrices A and B. Thus, a broad variety of linear hypotheses on the parameter $\Pi = \alpha\beta'$ can be investigated, notably inference on the coupling directions and the effect of system disequilibrium on individual oscillators.

Evaluation of the hypotheses H_α, H_β, and $H_{\alpha,\beta}$ all lead to similar likelihood ratio tests. To calculate the test statistic, solve again the eigenvalue problem (I.49) for the

unrestricted model, and dependent on the restrictions A and/or B obtain eigenvalues λ_i^* for the restricted model. The LRT statistic is then given by

$$-2\log Q\big(H_0|H_r\big) = T\sum_{i=1}^{r}\log\Big(\frac{1-\lambda_i^*}{1-\hat{\lambda}_i}\Big), \tag{I.26}$$

where H_0 is a generic substitute for any of $H_\alpha, H_\beta, H_{\alpha,\beta}$. Each of these statistics has an asymptotic χ^2 distribution with varying degrees of freedom (df),

$$\begin{aligned}
-2\log Q\big(H_\alpha|H(p)\big) &\text{ has } r(p-m) \text{ df}\\
-2\log Q\big(H_\beta|H(p)\big) &\text{ has } r(p-s) \text{ df}\\
-2\log Q\big(H_{\alpha,\beta}|H(p)\big) &\text{ has } r(p-m)+r(p-s) \text{ df},
\end{aligned}$$

where m and s are the column dimensions of the matrices A and B, respectively. This shows that once $\operatorname{rank}(\Pi)$ is determined, statistical inference for α and β can be carried out, relatively straightforward. As for the rank determination, an alternative to the χ^2 approximation for inference on α and β is to perform bootstrapping for the test (I.26), see Boswijk et al. (2016).

4 Numerical simulations

4.1 General setup

We perform a series of experiments with a system of $p = 3$ linearly coupled Winfree oscillators such that $z_t \in \mathbb{R}^6$ and $f(\phi_t) = \alpha\beta'\phi_t$. Hence, for $i = 1, 2, 3$, we have a DGP with

$$g(z_t)_i = f(\phi_t)_i = (\alpha\beta'\phi_t)_i = \alpha_i\sum_{j=1}^{3}\beta_j\phi_{jt}. \tag{I.27}$$

Since we examine simulations from the Winfree oscillator, our cointegration model will be misspecified since the amplitude is not deterministic and linear, but rather stochastic and fluctuating. However, since the amplitude of the Winfree oscillator has a relatively steady level (of course this also depends on the noise level), due to the squared multiplicative term in the amplitude process, we can approximate it as a constant. Hence we will do so in terms of analyzing the phase process as a cointegrating system. This also implies in terms of parameter estimation for the phase process, the estimate of the constant μ is a pseudo estimate of the κ parameter for the amplitude process, and hence we will compare the estimates to the true value of κ.

For each experiment we simulate 1.000.000 iterations of the oscillator (I.10) using the Euler-Maruyama scheme with timestep $\widetilde{\Delta}t = 0.0002$ and then subsample for $\Delta t = 0.1$, thus obtaining $N = 2000$ (equidistant and discrete) observations of z_t for $t \in [0, 200)$. Subsampling every 5000th values diminishes the discretization error of the simulation scheme.

We use the same initial conditions,

$$z_0 = (1, 0, 0, 1, -1, 0)',$$

and parameters

$$\begin{aligned} \kappa &= (0.75, 1, 1)' \\ \Sigma_\phi &= \text{diag}(1, 1, 1) \\ \Sigma_\gamma &= \text{diag}(0.1, 0.1, 0.1) \end{aligned} \tag{I.28}$$

for all the experiments so that the only varying parameter is the coupling structure.

Note that the κ parameter for ϕ_{2t} is set equal to ϕ_{3t} to obtain similar simulated outcomes for some experiments to investigate whether we can distinguish between interaction and independence between these two. We set the cointegration parameters for each experiment individually to impose different coupling structures, and will refer to the relevant model by it's $\Pi_k, k = 0, 1, 2, 3$ matrix, where k defines the model structure (see Fig. I.1).

The discrete time model fitted to the data is specified as

$$\Delta\phi_n = P\phi_{n-1} - \mu + \varepsilon_n, \tag{I.29}$$

where the estimate $\hat{P}$ is used to obtain $\hat{\Pi}$ through (I.20). The reported estimate for $\hat{\mu}$ is scaled by the timestep Δt. Note that μ is not time-dependent and hence this model will fit a constant parameter μ to a varying quantity γ_t and thus it is misspecified as mentioned above. Model (I.29) is estimated for all 4 systems of three oscillators and we report the parameters $\hat{\Pi}$ and $\hat{\mu}$ for each system. The latter is compared to κ, which is the level parameter of the γ_t process.

In addition to a cointegration analysis we apply the *mean phase coherence measure*, see Mormann et al. (2000), bilaterally to the *wrapped* phases (i.e., $\phi_{it} \in [0, 2\pi)$ for $i = 1, 2, 3$)

$$R(\phi_{it}, \phi_{jt}) = \left| \frac{1}{N} \sum_{n=1}^{N} e^{i(\phi_{i,t_n} - \phi_{j,t_n})} \right|, \tag{I.30}$$

as an initial measure of synchronization between the phases in the system. If $R \approx 1$ this implies synchronization ($R = 1$ means that oscillators are perfectly phase locked). On the contrary, $R \approx 0$ implies that the distribution of phase differences is approximately uniformly distributed on the unit circle. Note that the mean phase coherence measure is symmetrical, like correlation, and therefore it cannot reveal uni-directional coupling. In order to determine the significance of the R measures, we bootstrapped critical values for the hypothesis $R = 0$. Hence, these values are the same for all experiments and presented along with the measured R values. We compare the resulting value of R to the conclusion of the cointegration analysis.

We use the same seed for all experiments so that the outcomes are fully comparable in terms of stochastic noise dW. First we run a simulation with uncoupled oscillators as a benchmark, and then continue with coupled systems as presented in Fig. I.1. Fig. I.2 display the x-coordinates for $t \in [100, 150]$ from a simulation of these four systems.

The data analysis is carried out using the free software package R, R Core Team (2017). The source code for simulation and bootstrapping procedures are written

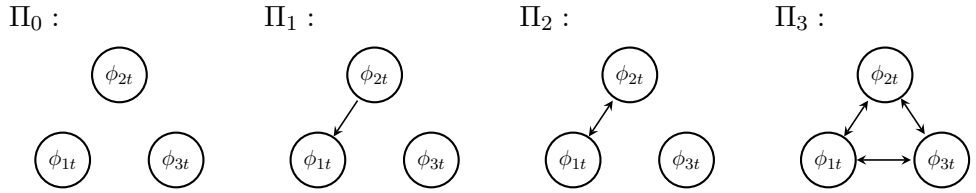

FIGURE I.1: Graphical representation of the four systems, represented by the $\Pi_i, i = 0, 1, 2, 3$ matrix. The arrows define the direction of interaction, hence $\phi_{2t} \to \phi_{1t}$ implies that ϕ_{2t} is influencing ϕ_{1t} (uni-directional coupling), and $\phi_{2t} \leftrightarrow \phi_{1t}$ denotes bi-directional coupling, i.e. ϕ_{1t}, ϕ_{2t} influence eachother.

in C++ to decrease the runtime, utilizing the interface package `Rcpp` for R and linear algebra package `RcppArmadillo` for C++. The source code is available in the package `cods` as supplementary material.

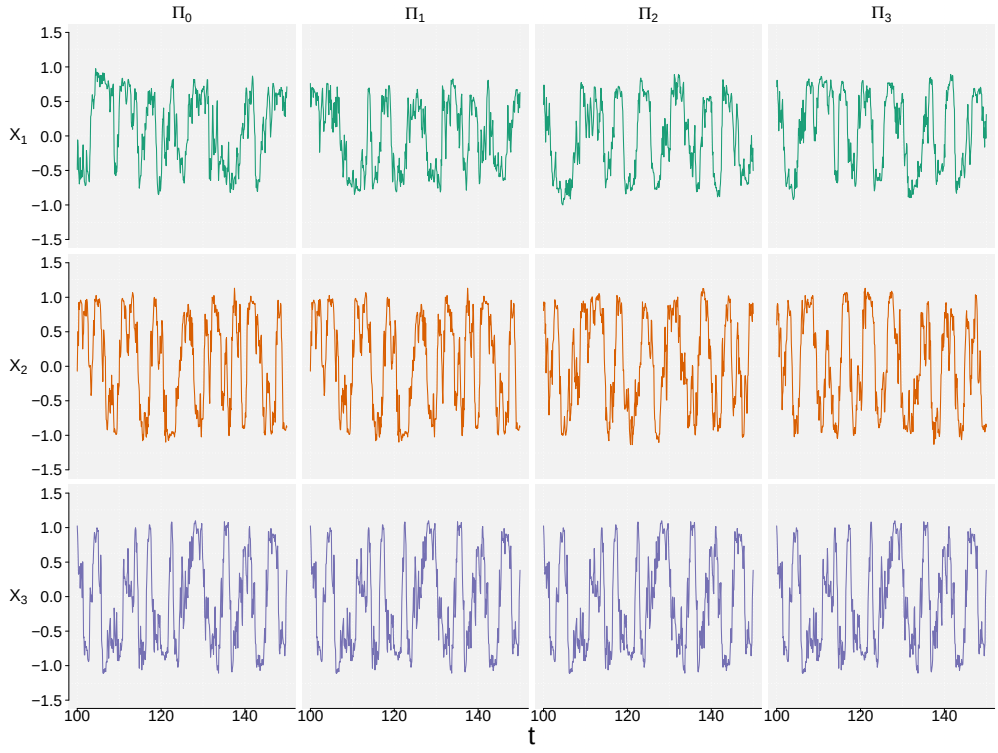

FIGURE I.2: 50 observations (x-coordinates only) from numerical simulation of the Winfree oscillator: the Π_0 column displays the independent model (I.31), the Π_1 column displays the uni-directional coupled model (I.32), the Π_2 column displays the bi-directional coupled model (I.33) and the Π_3 column displays the fully coupled model (I.34).

4.2 Independent oscillators

This experiment is used as a reference example. We set

$$\Pi_0 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{I.31}$$

so $\text{rank}(\Pi_0) = 0$ and there is no interaction in the system.

Simulating the model and unwrapping the phases, we obtain the top-left plot of Fig. I.3. Visual inspection of the plot could lead to the conclusion that ϕ_{2t} and ϕ_{3t}

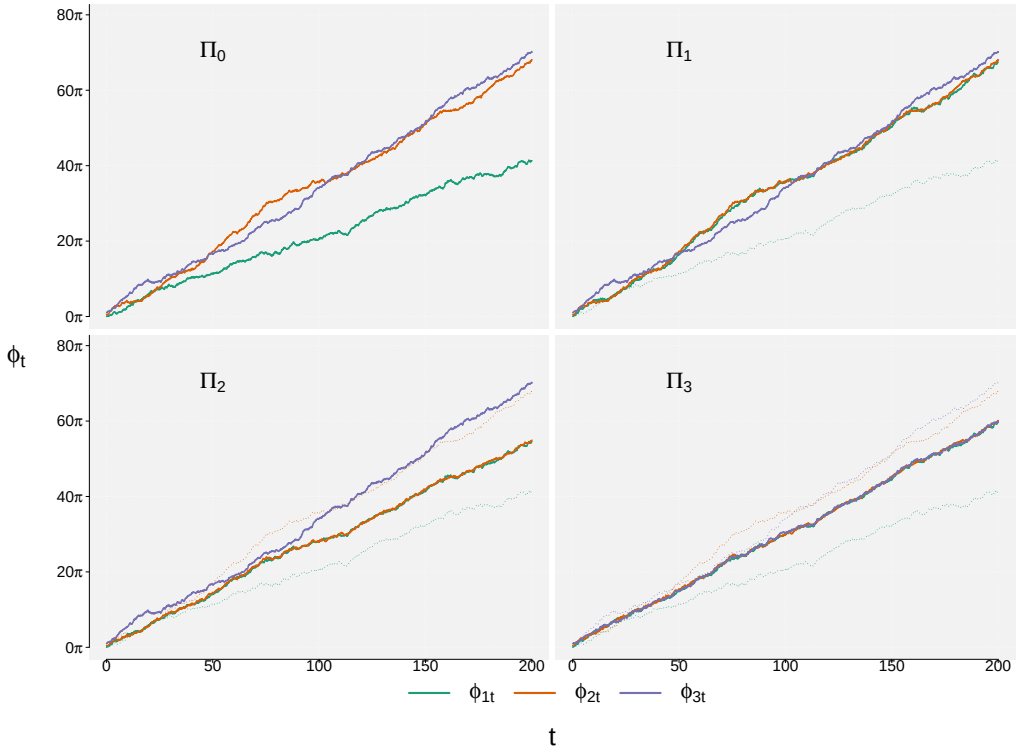

FIGURE I.3: Unwrapped phase processes from numerical simulation of the Winfree oscillator: Π_0 displays the independent model (I.31), Π_1 displays the uni-directional coupled model (I.32), Π_2 displays the bi-directional coupled model (I.33) and Π_3 displays the fully coupled model (I.34). The dotted lines represent the corresponding phases from the independent model in Π_0.

are coupled, however the mean phase coherence measure R for the phases indicates that this is not the case.

$$\begin{aligned} R(\phi_{1t}, \phi_{2t}) &= 0.025 \quad (0.170) \\ R(\phi_{1t}, \phi_{3t}) &= 0.073 \quad (0.168) \\ R(\phi_{2t}, \phi_{3t}) &= 0.078 \quad (0.176). \end{aligned}$$

The distribution of the mean phase coherence measure is unknown, but can be approximated by bootstrapping for $H_0 : R = 0$, that is for no synchronization present. 1000 bootstrap samples yield the reported 5% critical values in parentheses above ≈ 0.17, thus the mean phase coherence measure suggest no synchronization present, which is the case.

Performing now a rank test for the rank of Π_0 in the system, we obtain the first part of Table I.1.

The test does not reject the hypothesis $H_r : r = 0$, thus suggesting that there is no cointegration present in the system. This in turn implies that the oscillators are

Model	H_r	Test values	p-value
Π_0	$\mathbf{r = 0}$	$\mathbf{16.47}$	$\mathbf{0.663}$
	$r \leq 1$	3.94	0.753
	$r \leq 2$	0.05	0.812
Π_1	$r = 0$	118.39	0.000
	$\mathbf{r \leq 1}$	$\mathbf{4.30}$	$\mathbf{0.568}$
	$r \leq 2$	0.00	0.958
Π_2	$r = 0$	104.48	0.000
	$\mathbf{r \leq 1}$	$\mathbf{3.84}$	$\mathbf{0.707}$
	$r \leq 2$	0.03	0.843
Π_3	$r = 0$	157.81	0.000
	$r \leq 1$	63.82	0.000
	$\mathbf{r \leq 2}$	$\mathbf{0.00}$	$\mathbf{0.947}$

TABLE I.1: Rank tests for models $\Pi_i, i = 0, 1, 2, 3$ with the selected models indicated in bold. The test values are given by eq. (I.22) and p-values are determined by bootstrapping.

independent in terms of synchronization, in accordance with the DGP for Π_0, and with the mean phase coherence measure.

4.3 Uni-directional coupling

In this experiment we analyze a system with a uni-directional coupling. Let

$$\begin{aligned} \alpha &= (-0.5, 0, 0)' \\ \beta &= (1, -1, 0)' \end{aligned} \tag{I.32}$$

such that $\text{rank}(\Pi_1) = \text{rank}(\alpha\beta') = 1$, and we have the stationary relation $\phi_{1t} - \phi_{2t}$. Since $\alpha_2 = \alpha_3 = 0$, then ϕ_{2t} and ϕ_{3t} are acting independently, whereas ϕ_{1t} is influenced by ϕ_{2t}. Hence, the only coupling is $\phi_{2t} \to \phi_{1t}$.

The unwrapped phases for the simulation of model Π_1 are seen in the top-right of Fig. I.3. The dashed lines indicate the independent phases from the top-left of Fig. I.3, and we see that phases ϕ_{2t}, ϕ_{3t} are equal to their independent versions, whereas we now clearly find that ϕ_{1t} is attracted towards ϕ_{2t} due to the coupling structure in the system.

Examining the mean phase coherence in equation (I.30) for the system (note that $R(\phi_{2t}, \phi_{3t})$ is equal to the value in the previous section),

$$\begin{aligned} R(\phi_{1t}, \phi_{2t}) &= 0.321 \quad (0.170) \\ R(\phi_{1t}, \phi_{3t}) &= 0.049 \quad (0.168) \\ R(\phi_{2t}, \phi_{3t}) &= 0.078 \quad (0.176), \end{aligned}$$

we find indications of some synchronization between the phases ϕ_{1t} and ϕ_{2t} in the system compared to $R(\phi_{1t}, \phi_{2t})$ in the independent model. The value is significant on a 5% level as seen by the reported critical values, whereas for $R(\phi_{1t}, \phi_{3t})$ and $R(\phi_{2t}, \phi_{3t})$ the reported values are not. However, the mean phase coherence measure

Parameter	True value	Unrestricted estimates			Restricted α and β		
		Estimate	Std. Error	p value	Estimate	Std. Error	p value
α_1	-0.5	-0.527	0.049	< 0.001	-0.514	0.048	< 0.001
α_2	0	-0.050	0.049	0.307	0		
α_3	0	0.059	0.048	0.223	0		
β_1	1	1			1		
β_2	-1	-0.981			-1		
β_3	0	-0.016			0		
κ_1	0.75	0.765	0.076	< 0.001	0.638	0.081	< 0.001
κ_2	1	1.035	0.075	< 0.001	1.063	0.080	< 0.001
κ_3	1	1.119	0.074	< 0.001	1.086	0.080	< 0.001

TABLE I.2: Fitted model Π_1

does not recognize the uni-directional coupling as is the case here. Thus, it cannot distinguish between $\phi_{1t} \to \phi_{2t}$, $\phi_{1t} \leftarrow \phi_{2t}$ and $\phi_{1t} \leftrightarrow \phi_{2t}$.

Results from the rank test are in the second part of Table I.1. Here we see that $r = \text{rank}(\Pi_1) = 0$ is clearly rejected, whereas $r = 1$ cannot be rejected with a p-value of 0.568. This indicates the presence of a single cointegration relation, in accordance with the construction of the model.

Fitting the model with $r = 1$, we obtain the unrestricted MLE regression estimates in Table I.2. The cointegration relations are close to their true values (approximately within 1 standard error), and both α_2 and α_3 are statistically insignificant. Moreover, the estimates of β suggests a $1:1$ coupling between ϕ_1 and ϕ_2. Therefore, we perform a likelihood test for reducing the unrestricted model, with restrictions for both α and β

$$\begin{aligned} H_{\alpha,\beta} : \alpha &= A\psi, \text{ with } A = (1,0,0)' \\ \beta &= B\xi, \text{ with } B = (1,-1,0)', \end{aligned}$$

so that A fix $\alpha_2 = \alpha_3 = 0$ and B restricts to a 1:1 coupling. This yields the test statistic 3.617 which is χ^2 distributed with 4 degrees of freedom and hence implies a p-value of 0.460. Thus, we recover the true uni-directional coupling structure of the simulated phases. The fitted model is presented in the right of Table I.2.

The conclusion is that we have successfully identified the coupling structure of uni-directional coupled phases in a three dimensional system, with two independent phases, and one dependent. Since ϕ_{3t} is completely independent of ϕ_{1t} and ϕ_{2t} and $r = 1$ we can discard ϕ_{3t} when interpreting the cointegration in the system. Then we can interpret the cointegration parameter α as the coupling strength and β as the coupling scheme, here 1:1. If we had analyzed different data, with a β estimate close to $\hat{\beta} = (1,-n,0)'$, we could then identify a $n:1$ coupling between ϕ_{1t} and ϕ_{2t}. This can be seen from the fact that in this case $\alpha_k(\phi_{1t} - n\phi_{2t})$ would be a stationary relation, and thus ϕ_{2t} would rotate $\approx n$ times slower than ϕ_{1t}.

4.4 A bi-directional coupling with one independent oscillator

We now look at a system with

$$\begin{aligned}\alpha &= (-0.5, 0.5, 0)' \\ \beta &= (1, -1, 0)'.\end{aligned} \tag{I.33}$$

Hence, $\text{rank}(\Pi_2) = 1$ and again we have 1 stationary relation $\phi_{1t} - \phi_{2t}$, but now with only ϕ_{3t} independent, and a bidirectional coupling $\phi_{1t} \leftrightarrow \phi_{2t}$.

Simulating the Π_2-model we obtain the bottom-left of Fig. I.3. We have included the dashed lines again, as references for the independent system. If we contrast the bottom-left of Fig. I.3 with the top-right Fig. I.3, we now find that ϕ_{1t} and ϕ_{2t} are attracting each other, and hence they are both different from their independent versions. Since $|\alpha_1| = |\alpha_2|$, their coupling strength is equal, and the coupled phases lies roughly in the middle between the independent ones. If we look at the mean phase coherence measure for the pairwise interactions,

$$\begin{aligned}R(\phi_{1t}, \phi_{2t}) &= 0.590 \quad (0.170) \\ R(\phi_{1t}, \phi_{3t}) &= 0.144 \quad (0.168) \\ R(\phi_{2t}, \phi_{3t}) &= 0.126 \quad (0.176),\end{aligned}$$

we find relatively strong evidence of a coupling between the phases ϕ_{1t} and ϕ_{2t}, the value is higher than in the uni-directional case and it is (again) significant given the bootstrapped critical values. However, again we cannot distinguish between types of coupling structures.

Performing a rank test for cointegration in the system with Π_2, we see in the third part of Table I.1 that $H_r : r = 0$ is clearly rejected, and we find that the rank of Π_2 is estimated to 1 with a p-value of 0.707. Hence, we recover the correct dimension of the column space of β, and fitting a model with $r = 1$ yields the parameters in the left of Table I.3.

The only insignificant parameter for the model is α_3, which is in accordance with the construction of the Π_2 model. Specifying the hypothesis

$$\begin{aligned}H_{\alpha,\beta} : \alpha &= A\psi, \text{ with } A = \begin{pmatrix}1 & 0 \\ 0 & 1 \\ 0 & 0\end{pmatrix} \\ \beta &= B\xi, \text{ with } B = (1, -1, 0)',\end{aligned}$$

and performing a likelihood ratio test for the reduction yields a test statistic of 3.340, which follows a χ^2 with 3 degrees of freedom, and result in a p-value of 0.342. The fitted model is given in the middle of Table I.3. If we instead of $H_{\alpha,\beta}$ specify

$$\begin{aligned}H^*_{\alpha,\beta} : \alpha &= A\psi, \text{ with } A = (1, -1, 0)' \\ \beta &= B\xi, \text{ with } B = (1, -1, 0)',\end{aligned}$$

implying that $\alpha_1 = -\alpha_2$, we obtain a test statistic of 3.880, with 4 degrees of freedom, and a p-value of 0.423. Thus, we can also restrict the model to one where the coupling strengths are equal in magnitude. The fitted model is presented in the right part of Table I.3.

Para-meter	True value	Unrestricted estimates			Restricted α and β			Restricted α and β with A^*		
		Estimate	Std. Error	p value	Estimate	Std. Error	p value	Estimate	Std. Error	p value
α_1	-0.5	-0.530	0.071	< 0.001	-0.506	0.069	< 0.001	-0.475	0.069	< 0.001
α_2	0.5	0.450	0.069	< 0.001	0.443	0.067	< 0.001	0.475	0.067	< 0.001
α_3	0	0.087	0.070	0.214	0			0		
β_1	1	1			1			1		
β_2	-1	-0.970			-1			-1		
β_3	0	-0.022			0			0		
κ_1	0.75	0.754	0.072	< 0.001	0.646	0.076	< 0.001	0.660	0.076	< 0.001
κ_2	1	0.943	0.070	< 0.001	1.040	0.074	< 0.001	1.053	0.074	< 0.001
κ_3	1	1.103	0.071	< 0.001	1.086	0.075	< 0.001	1.086	0.075	< 0.001

TABLE I.3: Fitted model Π_2

Summing up, in a system of bi-directional coupled oscillators plus one independent, we can identify the correct coupling between them, including identifying the proportionally equal coupling strength between the coupled phases. Again we identify $r = 1$, and hence we can interpret the cointegration parameters as before, hence α is the coupling strength, and β the interaction, again 1:1 coupling.

4.5 Fully coupled system

We specify a system with full interaction between all phases.

$$\begin{aligned} \alpha &= \begin{pmatrix} -0.50 & 0.25 \\ 0.25 & -0.50 \\ 0.25 & 0.25 \end{pmatrix} \\ \beta &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix}. \end{aligned} \tag{I.34}$$

The α and β matrix are chosen, such that

$$\Pi_3 = \alpha\beta' = \begin{pmatrix} -0.50 & 0.25 & 0.25 \\ 0.25 & -0.50 & 0.25 \\ 0.25 & 0.25 & -0.50 \end{pmatrix}$$

inspired by the simplistic linearization of the Kuramoto model, as presented in eq. (I.14). Note that $\mathrm{rank}(\Pi_3) = 2$.

The simulated phases are shown in the bottom-right of Fig. I.3. Comparing to the dashed (independent) versions, we now find that all phases are different from their independent versions. It appears as if ϕ_{2t}, ϕ_{3t} dominate the system, since ϕ_{1t} is attracted closer to their independent versions than otherwise, but it is also a two-against one ($\kappa_2 = \kappa_3 \neq \kappa_1$) scheme, and we roughly observe that ϕ_{1t} is attracted $2/3$ towards ϕ_{2t}, ϕ_{3t}, whereas ϕ_{2t}, ϕ_{3t} are attracted $1/3$ towards ϕ_{1t}. So by the construction of the system, this behavior seems natural. We find that the mean phase coherence measure

$$\begin{aligned} R(\phi_{1t}, \phi_{2t}) &= 0.487 \quad (0.170) \\ R(\phi_{1t}, \phi_{3t}) &= 0.574 \quad (0.168) \\ R(\phi_{2t}, \phi_{3t}) &= 0.488 \quad (0.176), \end{aligned}$$

indicates bilateral synchronization for all phases, and all values are significant. The rank test also gives clear evidence of cointegration and we identify $r = 2$, as seen in the bottom part of Table I.1, where both the hypotheses $r = 0$ and $r = 1$ are rejected. Fitting a model with $r = 2$ yields the left half of Table I.4. The estimated κ's are close to their respective values, whereas some of the α parameters deviate (more than 1 standard error) from their true values. If we inspect the estimated $\hat{\Pi}_3$

$$\hat{\Pi}_3 = \begin{pmatrix} -0.611 & 0.231 & 0.378 \\ 0.241 & -0.437 & 0.196 \\ 0.343 & 0.207 & -0.549 \end{pmatrix}$$

and compare with the true Π_3 it looks better. The row sums are close to zero as they should be, and the signs are correct. The proportional coupling strengths are off

Parameter	True value	Unrestricted estimates			Restricted α and β		
		Estimate	Std. Error	p value	Estimate	Std. Error	p value
α_{11}	-0.50	-0.584	0.075	< 0.001	-0.569	0.075	< 0.001
α_{21}	0.25	0.232	0.073	< 0.001	0.241	0.072	< 0.001
α_{31}	0.25	0.326	0.072	< 0.001	0.328	0.072	< 0.001
α_{12}	0.25	0.223	0.067	< 0.001	0.224	0.067	< 0.001
α_{22}	-0.50	-0.423	0.064	< 0.001	-0.423	0.064	< 0.001
α_{32}	0.25	0.201	0.064	< 0.001	0.199	0.064	< 0.001
β_{11}	1	1			1		
β_{21}	0	0			0		
β_{31}	-1	-0.997			-1		
β_{12}	0	0			0		
β_{22}	1	1			1		
β_{32}	-1	-0.999			-1		
κ_1	0.75	0.712	0.076	< 0.001	0.607	0.083	< 0.001
κ_2	1	1.054	0.074	< 0.001	1.061	0.080	< 0.001
κ_2	1	1.023	0.073	< 0.001	1.130	0.080	< 0.001

TABLE I.4: Fitted model Π_3

though, especially between ϕ_{1t}, ϕ_{3t}, but it seems that Π_3 is relatively well estimated considering the identification issues. Recall that we can determine the subspaces sp(α) and sp(β) for continuous time cointegration models, see Kessler and Rahbek (2004), but that we have problems regarding the scaling of Π (see Section (3.3)).

Inspired by the fitted values, we restrict both matrices α and β

$$H_{\alpha,\beta} : \alpha = A\psi, \text{ with } A = \begin{pmatrix} -0.50 & 0.25 \\ 0.25 & -0.50 \\ 0.25 & 0.25 \end{pmatrix}$$
$$\beta = B\xi, \text{ with } B = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix}$$

we find that the test statistic is 1.73, χ^2 distributed with 4 degrees of freedom, and thus a p-value of 0.785. Hence, we can reduce the model to one with restrictions that generates the true structure of Π. The estimated model parameters are presented in Table I.4, and the corresponding $\hat{\Pi}$ is

$$\hat{\Pi}_3^* = \begin{pmatrix} -0.595 & 0.232 & 0.363 \\ 0.250 & -0.438 & 0.188 \\ 0.345 & 0.205 & -0.550 \end{pmatrix}.$$

Concluding on the fully coupled system, we find that we can correctly identify the dimension of the cointegration relations. We can also determine the coupling structure as given by the parameters α and β. However, interpretation in this experiment is more informative in terms of $\hat{\Pi}$, since with $r \geq 2$, the interpretation of cointegration parameters is not as intuitive as in the case of $r = 1$. We obtain a Π estimate that is reminiscent of the true matrix, with the true directions of the coupling, and strengths somewhat close to the actual values. Thus, we can interpret the system as

fully coupled, in a simplistic (linear) Kuramoto type model.

4.6 Strength of coupling and identification of interaction

In this section, we compare the mean phase coherence measure to the cointegration analysis with respect to interactions in the system. More specifically, we look at how strong the coupling constants in Π must be in order for the two methods to conclude correctly on interaction in the system. We reuse the parameter settings (I.34) from the fully coupled experiment, but use a scaled Π matrix $\Pi \rightarrow \varepsilon\Pi$, for $\varepsilon \in [0, 1]$, where ε controls the coupling strength. The higher ε is, the stronger the coupling, and hence the attraction between phases. Note that $\varepsilon = 0$ corresponds to the model Π_0 and $\varepsilon = 1$ corresponds to Π_3. The p-values are calculated using bootstrapping as presented by Cavaliere, Rahbek, and Taylor (2012) to obtain an estimate of the asymptotic distribution of the trace test statistics.

The aim is to investigate the rank test for varying ε compared to identification of interaction in the system using the mean phase coherence measure. Since low values of ε implies weak interaction, the expectation is that both methods will produce doubtful results in a low value regime. From the previous experiment on the fully coupled oscillators, the mean phase coherence measure produced low values on identifying the interaction of the system, hence we expect that the rank test will outperform for low values of ε.

The experimental setup is 100 repetitions for each value of ε, and in each repetition perform 500 bootstrap samples to estimate the p-value for the hypotheses $H_r : r = 0, 1, 2$. Fig. I.4 presents the median p-values for the rank test and median mean phase coherence measures against ε. The top row of the figure shows the p-values for $H_r : r = 0, 1, 2$ respectively, and the bottom row shows the mean phase coherence (R) measures for pairs of ϕ_{1t}, ϕ_{2t} and ϕ_{3t}. The dotted lines indicate the $p = 0.05$ value, under which we reject the hypotehsis. For the mean phase coherence measure, the 95% significance level for the hypothesis $R = 0$ has been determined numerically using bootstrapping and is indicated by the dotted lines. If the R-measure falls below this line, independence cannot be rejected.

Seen in the top row of Fig. I.4, at least half the simulations reject $H_r : r = 0$ for $\varepsilon > 0.12$, and at least half the simulations reject $H_r : r = 1$ for $\varepsilon > 0.11$. The test does not reject $H_r : r = 2$ for around 88% of the simulated samples for any values of ε. Thus, for $\varepsilon > 0.11$, we can conclude that there is interaction present in the system, and in most of the simulations we also recognize the true $\mathrm{rank}(\Pi) = 2$.

If we turn to the bottom row of Fig. 4, where the mean phase coherence measures are shown, we find that half the simulations does not reject the hypothesis $R = 0$ for $\varepsilon < 0.34, 0.36$ and 0.35, respectively, for $R(\phi_{1t}, \phi_{2t})$,$R(\phi_{1t}, \phi_{3t})$ and $R(\phi_{2t}, \phi_{3t})$, thus clearly indicating an inferior detection of interaction for small values of ε equivalent to weak couplings.

Concluding on this experiment, we find that the rank test detects interaction in the system already at relatively weak coupling strengths. In contrast to this, the coupling must be significantly stronger for a sound conclusion on interaction in the system when using mean phase coherence as a measure of interaction. Furthermore, when detecting interaction in the system, the rank test is also very capable of identifying the true rank of the system, despite a misspecified model. Higher sample sizes will of course improve the inference results.

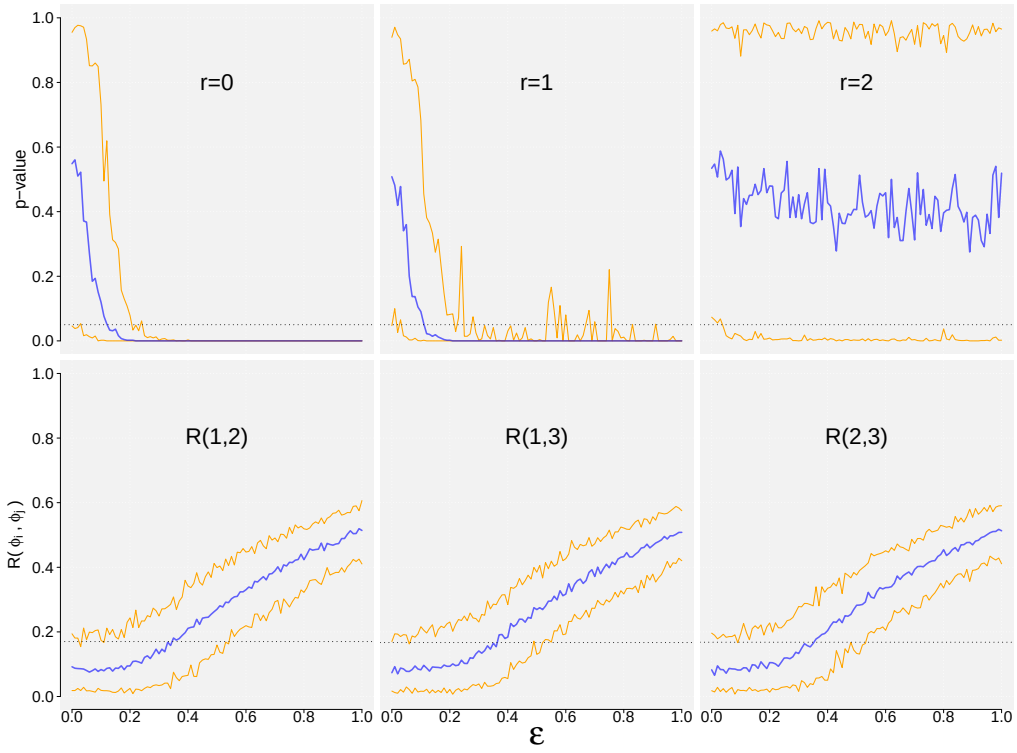

FIGURE I.4: Identification of interaction for varying coupling strengths ε for the model $\varepsilon\Pi_3$. Top row are hypotheses $H_r : r = 0, 1, 2$ using the trace test. The orange bounds indicates empirical quantiles ranging from 2.5% to 97.5% and the blue lines represents the median. The bottom row are mean phase coherence measures again with empirical quantiles illustrated. Dashed lines in the top row show the 5% confidence levels. Dashed lines in the bottom row show the 95% quantile for $H_0 : R(i, j) = 0$ at $\varepsilon = 0$, found by bootstrapping.

4.7 Consistency of the rank estimation

To investigate the consistency of the cointegration algorithm, we performed an experiment with 1000 repetitions of simulations for Winfree oscillators, the uni-directional coupling, the bi-directional and the fully coupled systems, respectively, and evaluating the rank test, using the same setup as in Section 4.1. Table I.5 present the percentages of conclusions regarding hypotheses $H_r : r = 0, r \leq 1, 2, 3$, for each model. Comparing with critical values at a 5% level, obtained by bootstrapping, see Cavaliere, Rahbek, and Taylor (2012), we find that comparing the percentage of simulations where the test correctly identifies the cointegration rank of 1 for uni- and bi-directional coupling are 76.8% and 69.8%, respectively, at a 5% significance level. For the fully coupled system the percentage is 85.5%, and for an independent system the percentage is 96.2%. These results show that identification of interaction in a system of coupled oscillators is quite precise, and the rank is underestimated in $\leq 2.5\%$ of the simulations for any model. In the case of independent or full interaction, the method is very good, whereas for systems with directed interaction, or interaction among some oscillators the frequency of overestimating the rank is $\approx 20 - 25\%$. This discrepancy seems intuitively correct, since for the latter systems the true model is a subset of the model of higher rank. As before higher sample sizes will of course improve the inference results.

Model	$r = 0$	$r \leq 1$	$r \leq 2$	$r \leq 3$
Independent	**96.2%**	2.2%	1.3%	0.3%
Uni-directional	1.7%	**76.8%**	19.0%	2.5%
Bi-directional	2.4%	**69.8%**	24.7%	3.1%
Fully coupled	0.0%	1.3%	**85.5%**	13.2%

TABLE I.5: Percentage of conclusions on $\mathrm{rank}(\Pi)$, at a 5% significance level for a sample size of 2000. Note that the conclusion $r \leq 3$ means that Π is of full rank and therefore invertible, hence $\beta = I_3$. Correct conclusions in **bold**.

Model	Rank test	$R(\phi_{1t}, \phi_{2t})$	$R(\phi_{1t}, \phi_{3t})$	$R(\phi_{2t}, \phi_{3t})$
Independent	3.8%	4.7%	4.4%	5.7%
Uni-directional	98.3%	99.8%	5.6%	4.4%
Bi-directional	97.6%	100.0%	7.2%	7.0%
Fully coupled	100%	100.0%	100%	100.0%

TABLE I.6: Percentage of conclusions on interaction indicated by the rank test and the mean phase coherence measures, at a 5% significance level for a sample size of 2000.

In Table I.6 we compare, in percentages, the conclusions on interaction in the systems, for each model. The values for the rank test presented here, are the summed values from Table I.5 for $r \neq 0$. We find that both methods are very adept in identifying interaction in these systems. The results, however, should be held against the previous section, where the rank test outperformed the mean phase coherence measure for weak coupling strength. Also noting the fact, that the mean phase coherence measure cannot account for uni-directional coupling, our overall conclusion is that in terms of identifying interaction in the system, the methods seem to perform equally well for stronger coupling, whereas in explaining the system architecture, a

cointegration analysis leaves us with more information on how the network is constructed.

5 Analysis of EEG data

Electroencephalography (EEG) signals are recordings from electrodes distributed on the scalp of subjects. The recorded brainwave patterns are, among others, used for diagnosing sleep disorders, coma or epilepsy. A study on 22 subjects experiencing epileptic seizures from the Children's Hospital Boston is presented by Shoeb (2009) with the aim of detecting seizures based on multiple hours of recordings for each individual. Fig. I.5 displays an EEG recording of a single subject during a period that include a seizure identified by Shoeb (2009) between 2996 seconds and 3036 seconds. The seizure is marked by two red dashed lines in Fig. I.5. The labels for the

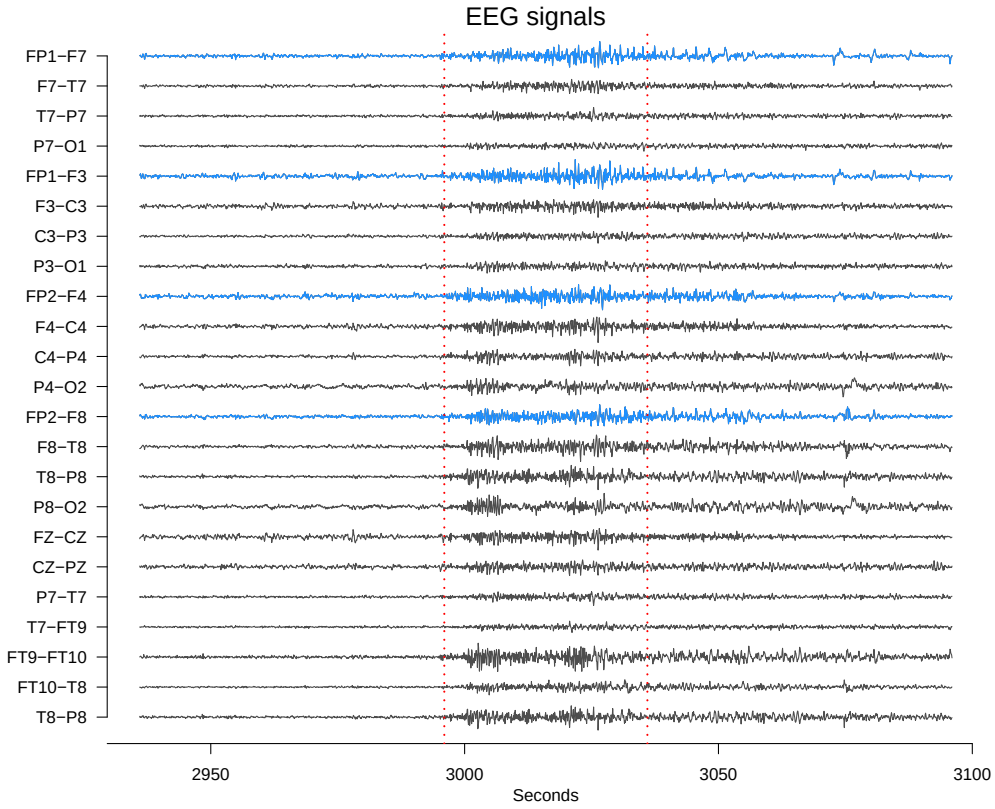

FIGURE I.5: EEG recording leading op to a seizure and afterwards for a 11 year old female subject. The interval [2996;3036] seconds, as indicated by the vertical red dashed lines, is defined by Shoeb (2009) as a seizure. We analyze the four blue signals, FP1-F7, FP1-F3, FP2-F4 and FP2-F8.

signals refer to the individual electrodes on the scalp. We analyze the four signals FP1-F7, FP1-F3, FP2-F4 and FP2-F8, where FP refer to the frontal lobes and F refer to a row of electrodes placed behind these. Even numbered electrodes are on the right side and odd numbered electrodes are on the left side. Smaller (larger) numberings imply that the electrode is placed closer to (further from) the center of the scalp. Hence FP1-F7, FP1-F3 are measurements from the left side, with F3 placed closer to the center than F7, and likewise for right side signals FP2-F4 and FP2-F8. The

	Prior to seizure	During seizure
$R_{\text{(FP1-F3; FP2-F4)}}$	0.480	0.542
$R_{\text{(FP1-F3; FP1-F7)}}$	0.535	0.644
$R_{\text{(FP1-F3; FP2-F8)}}$	0.295	0.184
$R_{\text{(FP2-F4; FP1-F7)}}$	0.321	0.350
$R_{\text{(FP2-F4; FP2-F8)}}$	0.486	0.342
$R_{\text{(FP1-F7; FP2-F8)}}$	0.525	0.379
Average	0.440	0.407

TABLE I.7: Mean Phase Coherence measures for EEG phases prior to and during the seizure.

electrodes for these four signals mirror each other on the left/right side of the scalp. We analyze the seizure period of 40 seconds and the 40 seconds leading up to the seizure, i.e. we analyze the two intervals [2956; 2996] and [2996; 3036] respectively, and refer to these as *prior to seizure* and *during seizure*. With a sample frequency of 256 measurements each second there are a total of 10240 measurements for each of the four signals during the 40 seconds intervals. For more details on the data, see Shoeb (2009). The objective is to compare two fitted cointegration models with interaction as in eq. (I.8) for each period:

$$d\phi_t = \alpha\beta' \begin{pmatrix} \phi_{t,\text{FP1-F3}} \\ \phi_{t,\text{FP1-F7}} \\ \phi_{t,\text{FP2-F4}} \\ \phi_{t,\text{FP2-F8}} \end{pmatrix} dt + \mu dt + \Sigma dW_t,$$

discretely observed for $t = 1, \ldots, 10240$ in each of the two intervals.

The phase processes of the four signals are estimated using the Hilbert transform (see section 2.3). Fig. I.6 shows the four signals in the two periods and their corresponding estimated unwrapped phase processes. Hence the offsets are in $[0, 2\pi)$ for the individual phase processes in each period. If we had not split the measurements at 2996 seconds, the phases in the bottom right of Fig. I.6 would be continuations of the phases in the bottom left. A visual inspection of Fig. I.6 shows that when transitioning to the seizure period, the phases change to a slower pace (the slopes decrease). Also, prior to the seizure all four phases are closer with no clear distinction between right side and left side phases. During the seizure, the phases split in two groups: right and left side respectively. This indicates that the model regime changes when transitioning into the seizure period. Table I.7 shows the mean phase coherence measures bilaterally for the 4 phase processes and the average of these. Comparing the columns we find no clear indication of a change in the phase regime when transitioning into the seizure period based on this measure, the average change is only 7.5%. However, the measure does indicate interaction in the system among all phases. Table I.8 displays the results of a rank test procedure for the system of the four EEG phase processes. In accordance with the indications from the mean phase coherence measure, the conclusion is a clear presence of cointegration during both periods. Prior to the seizure the rank test of $r \leq 2$ is close to the usual 5% significance level, hence the p-value here is determined using 5000 bootstrap samples, in contrast to the 2000 bootstrap samples used in the other interval, as the conclusion here is quite clear with a p-value ≈ 0.62. In both cases

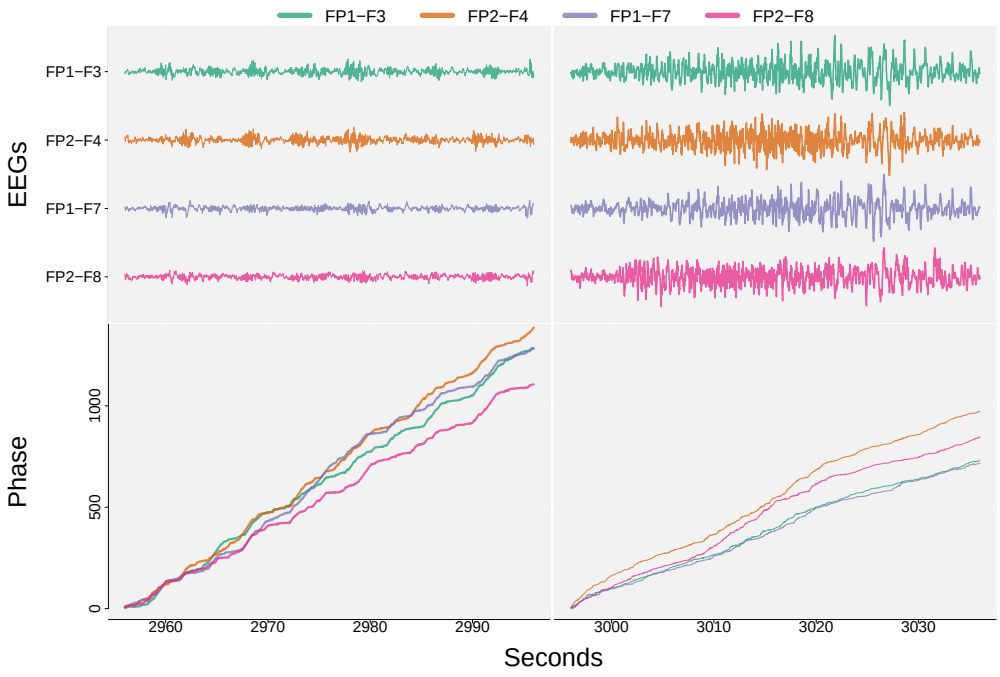

FIGURE I.6: FP1-F7, FP1-F3, FP2-F4 EEG signals and estimated phase processes for a 11 year old female subject. Top left: EEG signals prior to a seizure. Top right: EEG signals during a seizure. Bottom left: estimated phase processes prior to a seizure. Bottom right estimated phase processes during a seizure.

	Prior to seizure		During seizure	
H_r	Test values	p-value	Test values	p-value
$r = 0$	105.87	0.000	1132.64	0.000
$r \leq 1$	42.82	0.000	41.68	0.008
$\mathbf{r \leq 2}$	**9.98**	**0.053**	**7.19**	**0.618**
$r \leq 3$	0.46	0.439	0.72	0.786

TABLE I.8: Rank tests for EEG phases in the bottom of Fig. I.6. The rank is determined to $r = 2$ in both periods, although the conclusion is far stronger during the seizure. The significance of the statistics are found using 5000 bootstrap samples prior to the seizure due the border limit case of around 5%, during the seizure the p-value is determined from 2000 bootstrap samples.

Parameter	Prior to seizure			During seizure		
	Estimate	Std. Error	p value	Estimate	Std. Error	p value
$\alpha_{\text{FP1-F3},1}$	-0.100	0.018	< 0.001	-0.462	0.028	< 0.001
$\alpha_{\text{FP1-F7},1}$	-0.002	0.019	0.930	-0.308	0.032	< 0.001
$\alpha_{\text{FP2-F4},1}$	-0.035	0.017	0.044	-0.722	0.035	< 0.001
$\alpha_{\text{FP2-F8},1}$	-0.115	0.030	< 0.001	-0.648	0.042	< 0.001
$\alpha_{\text{FP1-F3},2}$	-0.117	0.016	< 0.001	0.041	0.033	0.212
$\alpha_{\text{FP1-F7},2}$	-0.024	0.016	0.147	0.071	0.037	0.057
$\alpha_{\text{FP2-F4},2}$	-0.026	0.015	0.084	0.173	0.041	< 0.001
$\alpha_{\text{FP2-F8},2}$	-0.049	0.026	0.063	0.468	0.049	< 0.001
$\beta_{\text{FP1-F3},1}$	1			1		
$\beta_{\text{FP1-F7},1}$	0			0		
$\beta_{\text{FP2-F4},1}$	-3.424			-0.036		
$\beta_{\text{FP2-F8},1}$	2.610			-0.573		
$\beta_{\text{FP1-F3},2}$	0			0		
$\beta_{\text{FP1-F7},2}$	1			1		
$\beta_{\text{FP2-F4},2}$	2.486			-0.840		
$\beta_{\text{FP2-F8},2}$	-3.631			0.188		
$\mu_{\text{FP1-F3}}$	25.210	2.162	< 0.001	39.647	1.307	< 0.001
$\mu_{\text{FP1-F7}}$	30.648	2.252	< 0.001	36.499	1.473	< 0.001
$\mu_{\text{FP2-F4}}$	39.058	2.107	< 0.001	58.268	1.608	< 0.001
$\mu_{\text{FP2-F8}}$	48.853	3.615	< 0.001	54.765	1.947	< 0.001

TABLE I.9: Fitted model for EEG phases F7-T7, T7-P7 and FP1-F7.

we choose the rank $r = 2$ for the system. The fitted models are presented in Table I.9 with the model fit prior to the seizure on the left side and the fit during the seizure on the right side. If we first note the estimated μ_i's, these are larger during the seizure and significantly so for FP1-F3 and FP2-F4, implying that these phase processes exhibit significantly higher intrinsic linear trends during the seizure. On the other hand, directly interpreting the cointegration parameters is not clear. Recall that these parameters specify subspaces, in this case within $\mathbb{R}^4$. We therefore look at the estimated $\hat{\Pi}$ matrices in Table I.10 to compare the models for each period. Here we can determine an all-to-all coupling during both periods and the estimated cointegration matrices show a clear difference for the two intervals. Prior to the seizure the right side signals FP2-F4 and FP2-F8 are much less influenced by the feedback in the system, whereas during the seizure both experience a much larger feedback from the left side signals FP1-F3 and FP1-F7 respectively. Surprisingly, the FP2-F8

	$\hat{\Pi}$ prior to seizure				$\hat{\Pi}$ during seizure			
	FP1-F3	FP1-F7	FP2-F4	FP2-F8	FP1-F3	FP1-F7	FP2-F4	FP2-F8
FP1-F3	4.388	1.572	-11.120	5.743	-5.305	-11.021	9.447	0.971
FP1-F7	1.519	0.892	-2.985	0.725	-4.335	-7.285	6.275	1.116
FP2-F4	0.540	-0.050	-1.971	1.589	-10.265	-17.047	14.686	2.681
FP2-F8	-0.733	-1.658	-1.613	4.108	-14.907	-14.729	12.909	5.776

TABLE I.10: Fitted $\hat{\Pi}$ matrices for the two periods. On the left side is the estimated matrix prior to the seizure, on the right side is the estimated matrix during the seizure.

signal does not seem to impose a large influence in the system in either interval. It is also interesting to note the changing signs in the two matrices. The two left side signals exhibit a positive feedback on themselves prior to the seizure, whereas during the seizure they impose a negative feedback both on themselves and the right side signals. This could possibly be part of an explanation of the slight kink seen in the phases around 3015-3020 seconds halfway through the seizure.

Concluding on this analysis we find, not surprisingly, a fully coupled 4 dimensional system with a clear change in the trends prior to and during the seizure. We find that during the seizure the interaction in the system is much stronger, suggesting the more distinctive phases shown in this interval. Including this temporal effect into a single cointegration model covering the full period by utilizing regime switching cointegration models, would be an interesting pursuit for future work.

6 Discussion

In this paper we have investigated the use of cointegration analysis to determine coupling structures in linearly phase coupled systems. Using these techniques we can with a good precision identify the coupling structure as a subspace for this type of model. A standard measure to identify synchronization in the literature is the mean phase coherence measure. Contrary to this standard measure, we can detect uni-directional coupling, and we can construct and test hypotheses on the model in form of linear restrictions in the estimated subspace. Furthermore, comparing the mean phase coherence measure with the cointegration analysis in Section 4.6, we found that cointegration detects interaction in a system more robustly and for weaker coupling strength than does the mean phase coherence measure. Combined with the fact that cointegration does not just provide a level of synchronization, but rather the structure of the synchronization mechanism, this technique can be used to infer system structures in a much more detailed manner. Of course this higher level of information comes at a cost, since the mean phase coherence measure is easily implemented for any system, whereas the cointegration analysis is more involved and time consuming.

Due to the linear nature of the cointegration theory used, we are not able to cover more complex models, such as the Kuramoto model. Thus, an important extension for future work would be to allow for nonlinear coupling functions. However, the linear structure appears naturally when considering a linearization around some phase-locked state, such as for systems showing synchrony or asynchrony. Another interesting pursuit is to extend the model framework to include nonlinear deterministic trends, such that also models like the FitzHugh-Nagumo or the van der Pol oscillator would be covered. The model considered in this paper was constructed from the starting point of the phase process in the spirit of the Kuramoto model, and noise was added on this level. Another approach would be to start from a biological model or a reduction thereof and introduce the noise on the DGP. This would also lead to non-linearities both in drift and diffusion of the phase process. Finally, high dimensional systems are a major challenge in the area of coupled oscillators, hence it would only be natural to investigate cointegration properties of high dimensional systems. A system of more than two synchronizing oscillators that are nonlinearly phase coupled, facilitate chaotic behavior since phases can then bilaterally attract and repel each other. When the number of oscillators increase, one quickly ends

up with intuitive shortcomings. The number of parameters rapidly increase with the dimension of the system, possibly leading to a desirable reduction to a sparse interaction structure. This is a key issue with the cointegration framework, which take into account all individual oscillators, as opposed to a mean-field approach that does not run into the same curse of dimensionality. The quality of the estimators will rapidly decrease with increasing dimension of the parameter space or numerical problems may arise. This problem might be alleviated by imposing a sparse interaction structure through a LASSO L_1 penalization.

Cointegration to identify coupling of oscillators has been attempted before in a neuroscience context by Dahlhaus and Neddermeyer (2012). There, the Kuramoto model is approximated for strongly phase coupled oscillators by setting $\sin(\phi_j - \phi_i) \approx \phi_j - \phi_i$, since the phase differences are assumed to be small. We have used the idea from Dahlhaus and Neddermeyer (2012) of analyzing the unwrapped multivariate phase process. Contrary to Dahlhaus and Neddermeyer (2012), however, we have not linearized the sine function to replicate Kuramoto, since this will cause a discrepancy when the phase difference of two oscillators is closer to π than 0 (or 2π). To mitigate this problem, we have instead taken the approach of designing a DGP with the properties we are interested in, and which allows for any phase differences. Furthermore, this DGP enables us to specify a cointegration model that comply with data from this DGP. Although it may not fully comply with a biological model, it can point to where necessary flexibility is needed in order to develop more realistic cointegration models for biological processes. A first attempt to analyze EEG signals with cointegration analysis with linear coupling structures has been presented. The results are promising, and reveal a finer dependence structure characterizing states of seizure and non-seizure in epileptic patients, which in this example was not possible from the simple Mean Phase Coherence measure. To fully explore the potential of the cointegration analysis for EEG signals, it would be useful to extend the model and analysis tools to allow for non-linearities and simultaneous treatment of many traces, as well as time varying coupling strengths.

Summing up, by applying cointegration as a technique to the field of coupled oscillators in biology, we open up for a whole new area of applications for this statistical theory. On the other hand, using cointegration methods, biologists can gain new insights into network structures, being able to fit models and carry out statistical hypothesis testing. If the cointegration framework presented in this paper can be extended to include the standard models currently used in the field, cointegration would prove a powerful analysis tool for researchers.

7 Acknowledgements

The work is part of the Dynamical Systems Interdisciplinary Network, University of Copenhagen.

A Derivation of an oscillating process with cointegrated phases

Using a transformation from polar to Cartesian coordinates coordinates, we can use Itô calculus to derive a data generating process $z_t = (x_{1t}, y_{1t}, \dots, x_{pt}, y_{pt})'$, which will yield a phase process, ϕ_t, and amplitude process, γ_t, that comply with (I.2) and (I.3).

Assume a system of p oscillators, such that $\phi_t \in \mathbb{R}^p$, $\gamma_t \in \mathbb{R}^p$ and $z_t \in \mathbb{R}^{2p}$. Let $\zeta_t = (\gamma_{1t}, \phi_{1t}, \dots, \gamma_{pt}, \phi_{pt})'$ denote the $2p$-dimensional process in the polar coordinates. For notational purposes, we omit the time index t for $z_t, \zeta_t, x_{kt}, y_{kt}, \phi_{kt}$ and γ_{kt}, where $k = 1, \dots, p$, then we have the following relations

$$z = (x_1, y_1, \dots, x_p, y_p)' = h(\zeta) = \begin{pmatrix} \gamma_1 \cos(\phi_1) \\ \gamma_1 \sin(\phi_1) \\ \vdots \\ \gamma_p \cos(\phi_p) \\ \gamma_p \sin(\phi_p) \end{pmatrix} \in \mathbb{R}^{2p}.$$

For $k = 1, \dots, p$, we find that the k'th coordinate pair $(x_k, y_k)'$ are given as the $2k-1$ and $2k$ entries in $z = h(\zeta)$, i.e., $(x_k, y_k)' = (z_{2k-1}, z_{2k})' = (h_{k-1}(\zeta), h_k(\zeta))' = (\gamma_k \cos(\phi_k), \gamma_k \sin(\phi_k))'$. Using the multivariate version of Itô's Lemma, we find that for $k = 1, \dots, p$

$$\begin{aligned}
dz_{2k-1} &= \sum_i \frac{\partial \gamma_k \cos(\phi_k)}{\partial \zeta_i} d\zeta_i + \frac{1}{2} \sum_{i,j} \frac{\partial^2 \gamma_k \cos(\phi_k)}{\partial \zeta_i \partial \zeta_j} d\zeta_i d\zeta_j \\
&= \frac{\partial \gamma_k \cos(\phi_k)}{\partial \gamma_k} d\gamma_k + \frac{\partial \gamma_k \cos(\phi_k)}{\partial \phi_k} d\phi_k \\
&\quad + \frac{1}{2} \left(\frac{\partial^2 \gamma_k \cos(\phi_k)}{\partial \gamma_k^2} (d\gamma_k)^2 + \frac{\partial^2 \gamma_k \cos(\phi_k)}{\partial \phi_k^2} (d\phi_k)^2 + 2 \frac{\partial^2 \gamma_k \cos(\phi_k)}{\partial \gamma_k \partial \phi_k} (d\gamma_k d\phi_k) \right) \\
&= \cos(\phi_k) d\gamma_k - \gamma_k \sin(\phi_k) d\phi_k - \frac{1}{2} \gamma_k \cos(\phi_k) (d\phi_k)^2 - \sin(\phi_k)(d\gamma_k d\phi_k) \\
&= x_k \gamma_k^{-1} d\gamma_k - y_k d\phi_k - \frac{1}{2} x_k (d\phi_k)^2 - y_k \gamma_k^{-1} d\gamma_k d\phi_k \\
dz_{2k} &= \sum_i \frac{\partial \gamma_k \sin(\phi_k)}{\partial \zeta_i} d\zeta_i + \frac{1}{2} \sum_{i,j} \frac{\partial^2 \gamma_k \sin(\phi_k)}{\partial \zeta_i \partial \zeta_j} d\zeta_i d\zeta_j \\
&= \dots \\
&= \sin(\phi_k) d\gamma_k + \gamma_k \cos(\phi_k) d\phi_k - \frac{1}{2} \gamma_k \sin(\phi_k) (d\phi_k)^2 + \cos(\phi_k)(d\gamma_k d\phi_k) \\
&= y_k \gamma_k^{-1} d\gamma_k + x_k d\phi_k - \frac{1}{2} y_k (d\phi_k)^2 + x_k \gamma_k^{-1} d\gamma_k d\phi_k.
\end{aligned}$$

Note that γ_k, ϕ_k are both uni-variate processes. Combining the expressions for x_k and y_k, we find

$$\begin{aligned}
d \begin{pmatrix} x_k \\ y_k \end{pmatrix} &= \begin{pmatrix} x_k \\ y_k \end{pmatrix} \gamma_k^{-1} d\gamma_k + \begin{pmatrix} -y_k \\ x_k \end{pmatrix} d\phi_k - \frac{1}{2} \begin{pmatrix} x_k \\ y_k \end{pmatrix} (d\phi_k)^2 + \begin{pmatrix} x_k \\ y_k \end{pmatrix} \gamma_k^{-1} d\gamma_k d\phi_k \\
&= \begin{pmatrix} (1 + d\phi_k)\gamma_k^{-1} d\gamma_k - \frac{1}{2}(d\phi_k)^2 & -d\phi_k \\ d\phi_k & (1 + d\phi_k)\gamma_k^{-1} d\gamma_k - \frac{1}{2}(d\phi_k)^2 \end{pmatrix} \begin{pmatrix} x_k \\ y_k \end{pmatrix}
\end{aligned} \tag{I.35}$$

Insert the relations

$$\begin{aligned} d\gamma_k &= g_k(\phi,\gamma)dt + \sigma_k^\gamma dW_k^\gamma \\ d\phi_k &= f_k(\phi,\gamma)dt + \sigma_k^\phi dW_k^\phi \\ (d\phi_k)^2 &= (\sigma_k^\phi)^2 dt \\ d\gamma_k d\phi_k &= \sigma_k^\gamma \sigma_k^\phi dt \\ \gamma_k &= \sqrt{x_k^2 + y_k^2} \end{aligned}$$

into (I.35) and obtain

$$\begin{aligned} d\begin{pmatrix} x_k \\ y_k \end{pmatrix} = &\underbrace{\begin{pmatrix} -\frac{1}{2}(\sigma_k^\phi)^2 & -f_k(\phi,\gamma) \\ f_k(\phi,\gamma) & -\frac{1}{2}(\sigma_k^\phi)^2 \end{pmatrix}\begin{pmatrix} x_k \\ y_k \end{pmatrix} dt + \begin{pmatrix} 0 & -\sigma_k^\phi \\ \sigma_k^\phi & 0 \end{pmatrix}\begin{pmatrix} x_k \\ y_k \end{pmatrix} dW_k^\phi}_{\text{"Phase related"}} \\ &\underbrace{+ \frac{g_k(\phi,\gamma) + \sigma_k^\gamma \sigma_k^\phi}{\sqrt{x_k^2 + y_k^2}}\begin{pmatrix} x_k \\ y_k \end{pmatrix} dt + \frac{\sigma_k^\gamma}{\sqrt{x_k^2 + y_k^2}}\begin{pmatrix} x_k \\ y_k \end{pmatrix} dW_k^\gamma}_{\text{"Amplitude related"}} \end{aligned} \tag{I.36}$$

The quotation marks in the description in (I.36) imply that one can intuitively interpret the system in this way, but the system is not mathematically split into these parts, as clearly f and g both depends on ϕ and γ, and σ_k^ϕ enters in the "amplitude" part.

Generalizing (I.36) we find

$$dz_t = (R_t + Q_t)z_t dt + a(z_t, \Sigma_\phi)dW^\phi + b(z_t, \Sigma_\gamma)dW^\gamma, \tag{I.37}$$

where R_t is a block diagonal matrix of 2×2 rotation matrices and Q_t is a diagonal matrix of amplitude dependent adjustments. The noise is composed of a sum of two state dependent multivariate processes, where the functions a, b define the noise as given in (I.36). The time index t has been added in (I.37) to emphasize the time dependency of the matrices R_t and Q_t.

B Rank test for Π and estimation of cointegrated models

Here we outline the procedure for determining the cointegration rank and estimating parameters in model (I.2). For a thorough presentation of this method, see Johansen (1996).

We refer to model (I.16) with $\text{rank}(P) = r$ and $\mu_t = \mu$ as H_r. Using this categorization, we have a nested sequence of models

$$H_0 \subset \cdots \subset H_r \subset \cdots \subset H_r,$$

which enables us to specify likelihood-ratio tests for the hypothesis H_r given H_{r+1} or H_r given H_p, where H_p is the unrestricted model. The first critical step for analyzing (I.2), is to determine the cointegration space dimension r. Given r, we can estimate the parameters in the model using *reduced rank regression* and ordinary least squares

(OLS). The first step is to remove the deterministic trend by regression, then reduced rank regression is used to estimate b by solving an eigenvalue problem, and finally the remaining parameters are estimated by OLS, using the estimated $\hat{b}$.

To describe the regression procedure, some convenient notation needs to be established. Define for $n = 1, \ldots, N$, the following functions of the data, $\Upsilon_{0t_n} = \Delta\phi_{t_n}, \Upsilon_{1t_n} = \phi_{t_{n-1}}$ and let $\varepsilon_{t_n} \sim \mathcal{N}_p(0, \Omega)$.

The log-likelihood function is then (up to a constant)

$$\log L(a, b, \mu, \Omega) = -\frac{1}{2} N \log|\Omega| - \frac{1}{2}\sum_{n=1}^{N} (\Upsilon_{0t_n} - ab'\Upsilon_{1t_n} - \mu)'\Omega^{-1}(\Upsilon_{0t_n} - ab'\Upsilon_{1t_n} - \mu). \tag{I.38}$$

Define

$$\begin{aligned} \mathbb{R}^{p\times p} \ni M_{ij} &= N^{-1}\sum_{n=1}^{N} \Upsilon_{it_n}\Upsilon'_{jt_n}, \text{ for } i, j = 0, 1 \\ \mathbb{R}^{p} \ni M_{i2} &= N^{-1}\sum_{n=1}^{N} \Upsilon_{it_n}, \text{ for } i = 0, 1 \\ \mathbb{R} \ni M_{22} &= 1. \end{aligned} \tag{I.39}$$

Note that $M_{ij} = M'_{ji}$. Then the estimate of μ given a and b is

$$\begin{aligned} \hat{\mu}(a, b) &= M_{02} - ab12 \\ &= N^{-1}\sum_{n=1}^{N} \left(\Delta\phi_{t_n} - ab'\phi_{t_{n-1}}\right) \end{aligned} \tag{I.40}$$

Define the residuals

$$\begin{aligned} R_{0t_n} &= \Upsilon_{0t_n} - M_{02} = \Delta\phi_{t_n} - N^{-1}\sum_{n=1}^{N} \Delta\phi_{t_n} \\ R_{1t_n} &= \Upsilon_{1t_n} - M_{12} = \phi_{t_{n-1}} - N^{-1}\sum_{n=1}^{N} \phi_{t_{n-1}} \end{aligned} \tag{I.41}$$

With these preliminary steps, we obtain the profiled likelihood function

$$\log L(a, b, \Omega) = -\frac{1}{2} N \log|\Omega| - \frac{1}{2}\sum_{n=1}^{N} (R_{0t_n} - ab'R_{1t_n})'\Omega^{-1}(R_{0t_n} - ab'R_{1t_n}), \tag{I.42}$$

equivalent to the regression equation

$$R_{0t_n} = ab'R_{1t_n} + \hat{\varepsilon}_{t_n}. \tag{I.43}$$

Equation (I.43) is estimated as a reduced rank regression, by solving for eigenvalues. Define

$$S_{ij} = N^{-1}\sum_{n=1}^{N} R_{it_n}R'_{jt_n} = M_{ij} - M_{i2}M_{2j}, \text{ for } i, j = 0, 1. \tag{I.44}$$

Then for a fixed b, we obtain estimates for a and Ω by OLS with $b'R_{1t_n}$ as the independent variable,

$$\hat{a}(b) = S_{01}b(b'S_{11}b)^{-1}, \tag{I.45}$$

$$\hat{\Omega}(b) = S_{00} - S_{01}b(b'S_{11}b)^{-1}b'S_{10}, \tag{I.46}$$

and the likelihood is then maximized as

$$L_{\max}^{-2/N}(b) = |S_{00} - S_{01}b(b'S_{11}b)^{-1}b'S_{10}|.$$

Using the Schur complement for the matrix

$$\begin{vmatrix} S_{00} & S_{01}b \\ b'S_{10} & b'S_{11}b \end{vmatrix}, \tag{I.47}$$

we find

$$|S_{00} - S_{01}b(b'S_{11}b)^{-1}b'S_{10}| = |S_{00}||b'(S_{11} - S_{10}S_{00}^{-1}S_{01})b|/|b'S_{11}b|. \tag{I.48}$$

Equation (I.48) is maximized for all $p \times r$ matrices by solving for the p eigenvalues λ_i in

$$|\lambda S_{11} - S_{10}S_{00}^{-1}S_{01}| = 0, \tag{I.49}$$

such that

$$\lambda_i S_{11} v_i = S_{10}S_{00}^{-1}S_{01}v_i,$$

and the $p \times 1$ eigenvectors $v_i, i = 1, \ldots, p$ are normalized,

$$v_j' S_{11} v_i = \begin{cases} 1, \text{ for } i = j \\ 0, \text{ for } i \neq j. \end{cases} \tag{I.50}$$

Then for a given r, $\hat{b}$ $(p \times r)$ is given by the r eigenvectors, $v_1, \ldots, v_r$, corresponding to the r largest eigenvalues $\hat{\lambda}_1 > \cdots > \hat{\lambda}_r$, and the maximum value of the likelihood function with this $\hat{b}$ is

$$L_{\max}^{-2/N} = |S_{00}| \prod_{i=1}^{r} (1 - \hat{\lambda}_i). \tag{I.51}$$

Since (I.51) holds for $r = 0, \ldots, p$, where for $r = 0$ we set $\text{sp}(\hat{b}) = \{0\}$ and for $r = p$ we set $\text{sp}(\hat{b}) = \mathbb{R}^p$, we have solved for all possible ranks r once and for all, and we can form the likelihood ratio test

$$-2 \log Q\big(H_r | H_p\big) = -N \sum_{i=r+1}^{p} \log(1 - \hat{\lambda}_i), \tag{I.52}$$

for the model H_r versus the model H_p. Equation (I.52) is known as the *trace statistic*, whereas

$$-2 \log Q\big(H_r | H_{r+1}\big) = -N \log(1 - \hat{\lambda}_{r+1}), \tag{I.53}$$

is known as the *maximum eigenvalue statistic*. The asymptotic distributions of (I.52) and (I.53) are both non-standard mixed Gaussian. Critical values for these can be found via simulation of a $p - r$ dimensional Brownian motion and using either the *trace* or *maximum eigenvalue* of a specially constructed $(p-r) \times (p-r)$ matrix, where the construction of the matrix depends on the structure of the deterministic terms in the model. Another possibility is to use *bootstrapping* as presented by Cavaliere, Rahbek, and Taylor (2012).

Using the trace test (I.52), the rank r is then determined by proceeding as follows

1. Initialize with $r = 0$.
2. For $r = 0, \ldots, p-1$, if H_r versus H_p is rejected, set $r \to r + 1$ and calculate (I.52).
3. Repeat step 2 until H_r versus H_p cannot be rejected, and set $\mathrm{rank}(\Pi) = r$.
4. If $r + 1 = p$, set $r = p$.

When the rank r is determined, then $\hat{b}$ is used for estimating the remaining parameters which follows from equations (I.40),(I.45) and (I.46).

Observe some conveniences of Johansens procedure. First, all the p eigenvalues are determined at the same time. Thus, the eigenvalue problem only needs to be solved once through the whole procedure. Secondly, with tabulated or simulated values for the likelihood ratio tests, determining r follows a simple procedure. Finally with b fixed, the remaining parameters are estimated using OLS.

MANUSCRIPT II

Cointegration analysis of high-dimensional linear Kuramoto networks

Jacob Østergaard, Anders Rahbek and Susanne Ditlevsen

Working paper.

1 Introduction

High-dimensional data is becoming ubiquitous in contemporary statistics and this development is also recognized in neuroscience where laboratory techniques facilitate recording an ever increasing number of neurons simultaneously. In Østergaard, Rahbek, and Ditlevsen (2017b) we explored a novel approach to infer network structure. By studying a system of coupled processes and using cointegration analysis to infer the coupling structure it was demonstrated how both uni-directional, bi-directional and all-to-all coupling was concluded based on observed oscillating processes. However, in Østergaard, Rahbek, and Ditlevsen (2017b) the studied systems only included three-dimensional processes which does not agree well with the interest in analysis tools for high-dimensional data series. In this paper we explore how cointegration analysis performs in high-dimensions in terms of both rank and parameter estimation. We assume a system of cointegrated processes with i.i.d. Gaussian innovations and perform likelihood estimation, given the rank. In regard to determining the rank, we demonstrate the performance of bootstrapping in a high-dimensional setting for simulated processes. We also define a linear version of the classic Kuramoto (1984) model which implies a symmetric design of the system. To include this in the estimation procedure, we define a new low-rank estimator of the system matrix, under the restriction of symmetry. This estimator diverge from the standard maximum likelihood estimation using reduced rank regression.

In the following, I_p denotes the p-dimensional identity matrix. For a real matrix $M \in \mathbb{R}^{p\times r}$, $M_\perp$ denote the orthogonal complement, such that $M_\perp' M = 0$. The matrix determinant operator is denoted $|\cdot|$. Also, an integrated process is implicitly assumed $I(1)$ (integrated of order 1) and $I(0)$ denote a process that can be given initial values such that it is stationary. Finally, time is assumed discrete (with positive index) throughout unless otherwise noted. Initial values are either assumed known or explicitly stated.

2 Cointegration

Assume the p-dimensional process

$$y_n = Ay_{n-1} + \mu + \varepsilon_n, \text{ for } n = 0, 1, \dots \tag{II.1}$$

where $A \in \mathbb{R}^{p\times p}$, $\mu \in \mathbb{R}^p$ and $\varepsilon_n \sim \mathcal{N}_p(0, \Omega)$. Assume also that (II.1) is integrated and let $\Pi = A - I_p$. The $I(1)$ assumption implies that $\operatorname{rank}(\Pi) = r < p$ such that $\Pi = \alpha\beta'$. If y_n satisfy the following criteria

- the characteristic polynomial for (II.1) has $p - r$ roots at $z = 1$ (unit roots) and all other roots lie outside the unit circle,
- the matrices $\alpha, \beta \in \mathbb{R}^{p\times r}$ have full column rank $r < p$,
- $|\alpha'_\perp \beta_\perp| \neq 0$,

then (II.1) is integrated of order 1. We therefore refer to the above criteria as the $I(1)$-conditions. These ensure that y_n is cointegrated with r cointegrating relations and $p - r$ stochastic (random walk type) trends. We will assume that $\mu = 0$, since deterministic trends are removed by regression to obtain a reduced rank regression model based on the residuals from this initial step. Hence, given N observations and writing (II.1) in the error correction form

$$\Delta y_n = \Pi y_{n-1} + \varepsilon_n, \text{ for } n = 0, 1, \dots, N, \tag{II.2}$$

where $\Delta y_n = y_n - y_{n-1}$, the log-likelihood function, omitting constant terms, is given by

$$\log L(\Pi, \Omega) = -\frac{N}{2}\log|\Omega| - \frac{1}{2}\sum_{n=1}^{N}(\Delta y_n - \Pi y_{n-1})\Omega^{-1}(\Delta y_n - y_{n-1})'. \tag{II.3}$$

Let H_r denote the hypothesis $\operatorname{rank}\Pi = r$. Then, under H_r, the maximum likelihood estimate of Π is given by a reduced rank regression (Johansen, 1996), for the decomposition $\Pi = \alpha\beta'$. Thus, under H_r, we denote the reduced rank estimator $\hat{\Pi} = \hat{\alpha}\hat{\beta}'$ as the unrestricted estimator with $\alpha, \beta \in \mathbb{R}^{p\times r}$.

Omitting the restriction to a reduced rank Π matrix for now, define the $p \times p$ moment matrices S_{ij}, $i, j = 0, 1$

$$\begin{aligned} S_{00} &= \frac{1}{N}\sum_{i=1}^{N}\Delta y_i(\Delta y_i)' \\ S_{01} &= \frac{1}{N}\sum_{i=1}^{N}\Delta y_i y'_{i-1} \\ S_{10} &= \frac{1}{N}\sum_{i=1}^{N} y_{i-1}(\Delta y_i)' \\ S_{11} &= \frac{1}{N}\sum_{i=1}^{N} y_{i-1}y'_{i-1}. \end{aligned} \tag{II.4}$$

The likelihood (II.3) can then be written with the $S_{ij}, i, j = 0, 1$ notation

$$\begin{aligned}\log L(\Pi, \Omega) &= -\frac{N}{2}\log|\Omega| - \frac{1}{2}\operatorname{Tr}\Big\{\frac{1}{N}\sum_{n=1}^{N}(\Delta y_n - \Pi y_{n-1})(\Delta y_n - \Pi y_{n-1})'\Omega^{-1}\Big\} \\ &= -\frac{N}{2}\log|\Omega| - \frac{1}{2}\operatorname{Tr}\Big\{\big(S_{00} - \Pi S_{10} - S_{01}\Pi' + \Pi S_{11}\Pi'\big)\Omega^{-1}\Big\}.\end{aligned}$$

2.1 Estimation under restriction to symmetry

Under H_r, Johansen (1996) derives a likelihood ratio test for linear restrictions on α, β. These are formulated as $\text{sp}(\alpha) \subset \text{sp}(A)$ and $\text{sp}(\beta) \subset \text{sp}(H)$, for given matrices A, H. However, in this paper we will concentrate on the restriction

$$\Pi \in \mathcal{S}_p = \{M \in \mathbb{R}^{p\times p} | M = M', \operatorname{rank} M = r\}, \tag{II.5}$$

i.e., Π must lie in S_p, the subset of symmetric $p \times p$ matrices with rank r. As such, the restrictions using matrices A, H from Johansen (1996) do not suffice for our purpose. The symmetry condition leads to the non-trivial problem of maximizing the likelihood of (II.2), under the two conditions $\operatorname{rank}(\Pi) = r$ and $\Pi = \Pi'$.

For the general case of approximating the $p \times p$ matrix Π with another, under the restriction of low rank and some given structure is known as *structured low rank approximation* (Chu, Funderlic, and Plemmons, 2003). For a symmetry restriction, the problem can be stated as

$$\min_M ||\Pi - M||_F, \text{ subject to } \operatorname{rank}(M) = r \text{ and } M \in \mathcal{S}_p \subset \mathbb{R}^{p\times p}, \tag{II.6}$$

where $||\cdot||_F$ denote the Frobenius norm and M is a generic matrix restricted to the structured subspace of real symmetric $p \times p$ matrices.

Under the rank condition only, Eckart and Young (1936) shows that the optimal solution to (II.6) is given by the singular value decomposition (SVD), a result which is known as the Eckart-Young-Mirsky theorem. Let $\Pi = U\Lambda V'$ denote the SVD of Π and assume that the diagonal entries of $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$ are ordered such that $\lambda_1 > \lambda_2 > \cdots > \lambda_p$. Partition U, V and Λ as

$$\begin{aligned}U &= (U_r, U_{p-r}) \\ V &= (V_r, V_{p-r}) \\ \Lambda &= \operatorname{diag}(\lambda_1, \ldots, \lambda_r, \lambda_{r+1}, \ldots, \lambda_p),\end{aligned}$$

such that $U_r, V_r \in \mathbb{R}^{p\times r}$ corresponds to the r largest entries of Λ and $U_{p-r}, V_{p-r} \in \mathbb{R}^{p\times(p-r)}$ corresponds to the last $p - r$ entries of Λ. Then, ignoring the symmetry constraint, the optimal solution to (II.6) is given by

$$\tilde{M} = U_r\Lambda_r V_r'. \tag{II.7}$$

However, once the symmetry is included the problem becomes much more difficult. Chu, Funderlic, and Plemmons (2003) notes that (II.6) may not have an optimal solution, whereas relaxing the rank constraint to $\operatorname{rank}(M) \leq r$ ensures a solution, given that the restricted subspace, which is S_p in this case, is non-empty.

Solving (II.6) can be done by numerical methods such as the *Lift-and-Project* algorithm:

1. Initialize $M^{(1)} = \Pi$. Then for $i = 1, 2, \ldots,$ do
2. *Lift*: compute a rank k approximation of $M^{(i)}$.
3. *Project*: find the projection $M^{(i)}$ of $M^{(i-1)}$ onto S_p, using $P(M) = \frac{M+M'}{2}$
4. Repeat steps 2 and 3 until convergence.

Another approach is penalized optimization:

$$\left|\left|\Pi - M + \kappa \sum_{ij} |M_{ij} - M_{ji}|\right|\right|_F, \kappa \in \mathbb{R}_+ \tag{II.8}$$

where $|\cdot|$ denotes the absolute value and M_{ij} refers to the ij'th element of M. For positive values of κ, non-symmetry is penalized and we obtain a near-symmetric approximation of M. Since the penalized optimization in (II.8) only optimize for the symmetry constraint, then in order to find a low-rank approximation of M, (II.7) must then be applied in turn after the penalized optimization, similar to the lift-and-project algorithm.

Opting for yet another approach, we first note that for the log-likelihood (II.3) the unrestricted OLS estimator for Π is found by solving for Π in the following matrix equation

$$-\partial\Pi S_{10} - S_{01}\partial\Pi' + \partial\Pi S_{11}\Pi' + \Pi S_{11}\partial\Pi' = 0, \tag{II.9}$$

which holds for

$$\hat{\Pi}_{\text{OLS}} = S_{01}S_{11}^{-1}. \tag{II.10}$$

For an arbitrary fixed Π, the covariance estimator is given as a function of Π

$$\Omega(\Pi) = S_{00} - \Pi S_{10} - S_{01}\Pi' + \Pi S_{11}\Pi'. \tag{II.11}$$

Inserting (II.11) in the likelihood, implies that the profile likelihood for Π, up to a constant, is given by

$$\begin{aligned}\log L(\Pi) &= -\frac{N}{2}\log|\Omega(\Pi)| \\ &= -\frac{N}{2}\log|S_{00} - \Pi S_{10} - S_{01}\Pi' + \Pi S_{11}\Pi'|.\end{aligned} \tag{II.12}$$

Note that if Π is symmetric and has reduced rank r it can be written as $\Pi = \beta\beta'$, for $\beta \in \mathbb{R}^{p\times r}$. Writing up the likelihood (II.12) and differentiating, then

$$\begin{aligned}\partial \log|\Omega(\partial\beta)| = \operatorname{Tr}\{\Omega(\beta)^{-1}(&-\partial\beta\beta'S_{10} - \beta\partial\beta'S_{10} - S_{01}\partial\beta\beta' - S_{01}\beta\partial\beta' \\ &+ \partial\beta\beta'S_{11}\beta\beta' + \beta\partial\beta'S_{11}\beta\beta' + \beta\beta'S_{11}\partial\beta\beta' + \beta\beta'S_{11}\beta\partial\beta')\} \\ = \operatorname{Tr}\{\Omega(\beta)^{-1}\big(&\ 2(\partial\beta\beta' + \beta\partial\beta')(S_{11}\beta\beta' - S_{10})\big)\}.\end{aligned}$$

Thus, in the case $\beta = 0$ (y_n is a p-dimensional Brownian motion), the score is 0. This suggest that deriving a symmetric estimator of Π through β might not be ideal.

Instead we can examine how the profile likelihood can be maximized under the restriction of $\Pi = \Pi'$. We focus on the symmetry constraint to obtain an estimator which can be approximated by a reduced rank version using the Eckart-Young-Mirsky theorem, since it does not seem possible to find a solution including both restrictions simultaneously. Maximizing (II.12) corresponds to minimizing the log determinant of $\log|\Omega(\Pi)|$

$$\max_{\Pi} \log L(\Pi) = \min_{\Pi} \log|\Omega(\Pi)|, \text{ subject to } \Pi = \Pi'$$

Considering the OLS estimator $\hat{\Pi}_{\text{OLS}} = S_{01}S_{11}^{-1}$, Fan and Hoffman (1955) prove that the nearest symmetric matrix to this, is the Hermitian part of $\hat{\Pi}_{\text{OLS}}$

$$\tilde{\Pi} = \frac{1}{2}(S_{01}S_{11}^{-1} + S_{11}^{-1}S_{10}) = \frac{1}{2}(\hat{\Pi}_{\text{OLS}} + \hat{\Pi}'_{\text{OLS}}), \tag{II.13}$$

in the sense that $||\hat{\Pi}_{\text{OLS}} - \tilde{\Pi}||_F \leq ||\hat{\Pi}_{\text{OLS}} - M||_F$, for any matrix $M \in \mathcal{S}_p$, with equality if and only if $M = \tilde{\Pi}$. Note that the Hermitian part of a $p \times p$ matrix corresponds to the projection onto $\mathcal{S}_p$. From this, we claim that (II.13) is the optimal choice of an otherwise unrestricted estimator of a symmetric Π, since $\hat{\Pi}_{\text{OLS}}$ is the optimal unrestricted estimator for Π and (II.13) is the optimal symmetric approximation of this estimator. These arguments are of course without any regard to the rank condition. Therefore, to obtain an estimate of rank r, we then invoke the SVD approximation from above to obtain

$$\hat{\Pi}_{\text{sym}} = U_r(\tilde{\Pi})\Lambda_r(\tilde{\Pi})(V_r'(\tilde{\Pi}), \tag{II.14}$$

where the notation $U_r(\cdot), \Lambda_r(\cdot), V_r(\cdot)$ refer to the SVD matrices of rank r with respect to the argument.

The maximized log-likehood is then given by (II.12), where

$$\hat{\Omega}(\hat{\Pi}_{\text{sym}}) = S_{00} - \hat{\Pi}_{\text{sym}}S_{10} - S_{01}\hat{\Pi}'_{\text{sym}} + \hat{\Pi}_{\text{sym}}S_{11}\hat{\Pi}'_{\text{sym}}.$$

3 Linear Kuramoto type system

The classic Kuramoto (1984) model defines a system of p coupled processes through the differential equations

$$\frac{d\theta_i}{dt} = \omega_i + \frac{K}{p}\sum_{j=1}^{p}\sin(\theta_j - \theta_i), \text{ for } i = 1, \ldots, p. \tag{II.15}$$

The variables θ_i are interpreted as the phases of limit cycle oscillators, with intrinsic frequencies ω_i and K denotes the strength of the coupling between the oscillators. By linearizing the sine function on $[0, \frac{\pi}{2})$ we can write the linearized Kuramoto model as

$$\frac{d\theta}{dt} = \omega + \frac{K}{p}\Pi\theta,$$

where $\theta = (\theta_1, \dots, \theta_p)'$, $\omega = (\omega_1, \dots, \omega_p)'$ and $\Pi \in \mathbb{R}^{p \times p}$ with the following simple structure

$$\Pi = \begin{pmatrix} p-1 & 1 & \dots & 1 \\ 1 & p-1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & p-1 \end{pmatrix} \tag{II.16}$$

and $\operatorname{rank}(\Pi) = p - 1$. Note that this linearization only holds when $\theta_j - \theta_i \leq \frac{\pi}{2}$ for all $i, j = 1, \dots, p$.

Consider now the following network structure inspired by the linearized version of (II.15), let Π denote a symmetric matrix with a block structure

$$\Pi = \begin{pmatrix} \Pi_1 & 0 & \dots & 0 \\ 0 & \Pi_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \Pi_k \end{pmatrix}, \tag{II.17}$$

where $\Pi_i \in \mathbb{R}^{p_i \times p_i}, i = 1, \dots, k$ are symmetric, reduced rank matrices with $\operatorname{rank}(\Pi_i) = r_i = p_i - 1, i = 1, \dots, k$. Hence, (II.17) implicitly defines a network of clusters. Note that the p_i's (and hence the r_i's) do not necessarily need to be identical. The network structure in (II.17) allows for cointegration since each Π_i has reduced rank, and $\operatorname{rank}(\Pi) = \sum_{i=1}^{k} \operatorname{rank}(\Pi_i) = \sum_{i=1}^{k} (p_i - 1)$. This implies that $\operatorname{rank}(\Pi) = p - k$. Interpreting this as a cointegrated system means that each cluster in (II.17) is driven by a stochastic $I(1)$ trend and hence the number of stochastic trends equals the number of independent clusters in the system.

The structure of Π in (II.17) implies that

$$\begin{aligned} \Pi &= \operatorname{diag}(\Pi_1, \dots, \Pi_k) \\ &= \operatorname{diag}(\alpha_1 \beta_1', \dots, \alpha_k \beta_k') \end{aligned}$$

where $\operatorname{diag}(\cdot)$ refers to block diagonalization. The rows/columns of Π must sum to 0, thus further restricting the possible structure of α, β. A possible construction of (II.17) from α, β with these restrictions, is to choose α, β such that

$$\alpha = \begin{pmatrix} \alpha_1 & 0 & \dots & 0 \\ 0 & \alpha_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \alpha_k \end{pmatrix} \tag{II.18}$$

$$\beta = \begin{pmatrix} \beta_1 & 0 & \dots & 0 \\ 0 & \beta_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \beta_k \end{pmatrix}, \tag{II.19}$$

where

$$\alpha_i = \begin{pmatrix} -r_i a_i & a_i & \dots & a_i \\ a_i & -r_i a_i & \dots & a_i \\ \vdots & \vdots & \ddots & \vdots \\ a_i & a_i & \dots & -r_i a_i \\ a_i & a_i & \dots & a_i \end{pmatrix} = a_i \begin{pmatrix} -r_i & 1 & \dots & 1 \\ 1 & -r_i & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & -r_i \\ 1 & 1 & \dots & 1 \end{pmatrix} \tag{II.20}$$

$$\beta_i = \begin{pmatrix} b_i & 0 & \dots & 0 \\ 0 & b_i & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & b_i \\ -b_i & -b_i & \dots & -b_i \end{pmatrix} = b_i \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ -1 & -1 & \dots & -1 \end{pmatrix}. \tag{II.21}$$

These choices of α_i, β_i implies that

$$\alpha_i \beta_i' = \begin{pmatrix} -r_i a_i b_i & a_i b_i & \dots & a_i b_i \\ a_i b_i & -r_i a_i b_i & \dots & a_i b_i \\ \vdots & \vdots & \ddots & \vdots \\ a_i b_i & a_i b_i & \dots & -r_i a_i b_i \end{pmatrix} = a_i b_i \begin{pmatrix} -r_i & 1 & \dots & 1 \\ 1 & -r_i & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & -r_i \end{pmatrix}, \tag{II.22}$$

which is symmetric, rows/columns sum to 0 (since it is $r_i \times r_i$) and couplings are equal within cluster i. Furthermore, note that within each cluster, the only two parameters are necessary to describe the interaction

1. $p_i = r_i + 1$, the size of the cluster, and
2. $c_i = a_i b_i$, the strength of coupling.

Each cluster i now corresponds to (II.16), where $\frac{K}{p} = a_i b_i$. Here a_i and b_i are not individually identifiable. As such, the usual normalization of β_i is imposed, such that the upper $r_i \times r_i$ matrix is the identity. The α_i matrix is adjusted accordingly, i.e. estimating the $a_i b_i$ level, rather than a_i by itself. We will refer to a cluster of size p_i as a p_i-cluster, hence a 2-cluster consist only of two coupled neurons. Note that this definition of a cluster does not exclude 1-clusters, since this is just a single independent neuron. In order to include this in the full α, β defintions (II.18) and (II.19), simply define $\alpha_i = \beta_i = 0$ when $p_i = 1$.

We now define a linear Kuramoto type system as a set of k clusters of sizes $p_i, i = 1, \dots, k$, such that the coupling matrix is on the form (II.17), where each cluster has a coupling structure equivalent to (II.22). This is the type of system which will be analyzed below.

4 High-dimensional estimation

Concerning the model (II.2), there are two crucial steps in obtaining estimates, as both the rank of the system as well as Π must be determined. As the dimension, p, of the model (II.2) increases to a high-dimensional setting, estimating $\hat{r}$ close to the true rank becomes progressively difficult as the rank test is prone to underestimate the rank. This is due to the construction of the test sequence, which starts with H_r against H_p for $r = 0$. If this is rejected, then r increases by 1 and this continues until H_r against H_p cannot be rejected or if $r + 1 = p$, such that $\hat{r} = p$. For some

true rank $r < p$, this sequence is bound to reach a conclusion. However, since H_r is evaluated in some distribution then, if the data is sufficient, it will eventually fall within a confidence region of this distribution, thus not rejecting H_r. This can occur prior to testing at the true rank r as shown in Figure II.1, for $N = 2000$ observations of a linear Kuramoto-type system as described above. Here the distribution of each test statistic was found by bootstrapping (Cavaliere, Rahbek, and Taylor, 2012). The

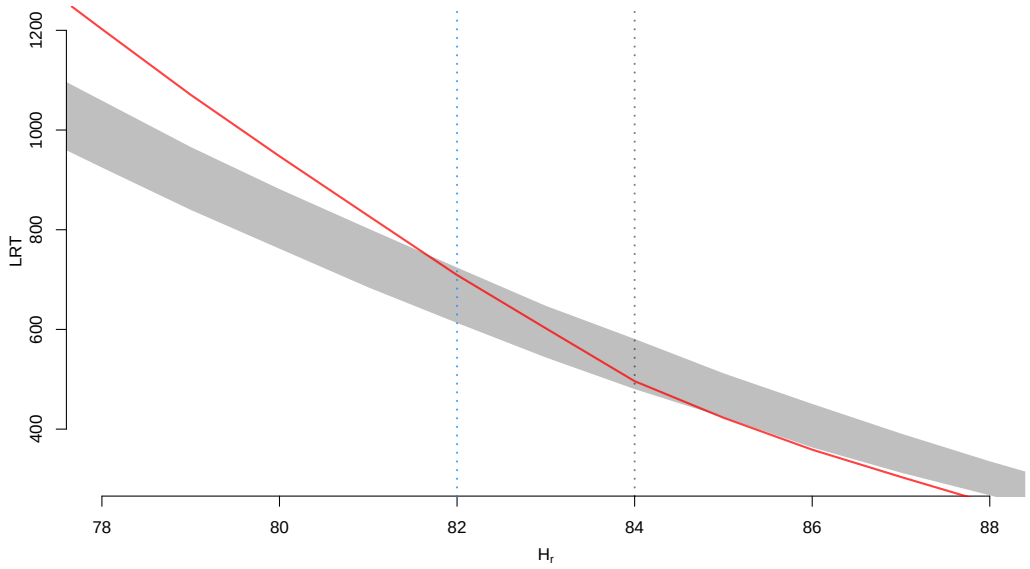

FIGURE II.1: Bootstrapped likelihood ratio test distribution of H_r versus H_p for $N = 2000$. 95% confidence bounds for the bootstrap test (gray) and the test values from the data (red). The test values enters the 95% region at $\hat{r} = 82$ (blue dashed line), whereas the true value is $r = 84$ (gray dashed line). The bootstrap test is based on 300 samples, increasing to 1000 does not change the conclusion of $\hat{r} = 82$.

figure demonstrates the 95% confidence bounds (gray) for the distribution of the likelihood ratio test under H_r, $r = 78, \ldots, 88$, using 300 bootstrap samples, when the true rank of the system is 84. The likelihood ratio test for the observed data (red) approaches the 95% confidence area from above, and is rejected up to, and including, $H_r : r = 81$. At $r = 82$, the test for the test falls within the 95% region and therefore it is not rejected. The bootstrap test therefore concludes that $\hat{r} = 82$ (blue dashed line), when in fact the true rank is $r = 84$ (gray dashed line). Inspecting Figure II.1 closely, the red curve displays a kink right at $r = 84$, indicating that at the true value there is a notable change in the likelihood ratio test trend. For a lower number of observations this is less visible, whereas for a higher number of observations, the kink is more pronounced. In the latter case, this implies that the numerical magnitude of the derivative of the likelihood ratio test curve, up to the true value of r, will increase with the number of observations and thus the test is more likely to find the true value of r, as the trend of the red curve will cut into the grey region at a larger angle. For the data used to produce Figure II.1, the conclusion was the same when using 1000 bootstrap samples, thus supporting the intuition that more observations will increase the accuracy, rather than just increasing the bootstrap sample, since this will only improve on the empirical distribution of the test, but not the trend of the test curve.

Following the discussion above on estimating the rank of a high dimensional system, it is of interest to asses how the estimates of (II.2) will perform, in regard to modeling

the true underlying structure of Π when the rank is underestimated. Intuitively we are approximating a subspace of dimension $\mathrm{rank}(\Pi) = r$ with a subspace of dimension $\mathrm{rank}(\hat{\Pi}) = \hat{r}$, thus whenever $\hat{r} \leq r$ then $\mathrm{sp}(\hat{\Pi}) \subseteq \mathrm{sp}(\Pi)$ and whenever $\hat{r} \geq r$ then $\mathrm{sp}(\Pi) \subseteq \mathrm{sp}(\hat{\Pi})$. Therefore, underestimating the rank implies that the basis of the estimation space will be a linear combination of combinations of the basis for the true Π-space. To see the effect of this, assume a linear Kuramoto-type system of just two clusters Π_1, Π_2, and define

$$\alpha_{10} = \begin{pmatrix} \alpha_1 \\ 0 \end{pmatrix} \qquad \alpha_{20} = \begin{pmatrix} 0 \\ \alpha_2 \end{pmatrix}$$
$$\beta_{10} = \begin{pmatrix} \beta_1 \\ 0 \end{pmatrix} \qquad \beta_{20} = \begin{pmatrix} 0 \\ \beta_2 \end{pmatrix}$$

such that $\alpha = (\alpha_{10}, \alpha_{20})$ and $\beta = (\beta_{10}, \beta_{20})$, then

$$\alpha\beta' = \alpha_{10}\beta_{10}' + \alpha_{20}\beta_{20}' = \begin{pmatrix} \alpha_1\beta_1' & 0 \\ 0 & \alpha_2\beta_2' \end{pmatrix}.$$

Assuming that the rank is underestimated, such that the estimated $\hat{\alpha}, \hat{\beta}$ are linear combinations of α_{i0} and β_{i0}, respectively,

$$\hat{\alpha} = \nu\alpha_{10} + (1-\nu)\alpha_{20}$$
$$\hat{\beta} = \nu\beta_{10} + (1-\nu)\beta_{20},$$

for $\nu \in [0, 1]$, then these are estimates in lower dimensional spaces than the true α, β. From this,

$$\begin{aligned}\hat{\alpha}\hat{\beta}' &= \nu^2\alpha_{10}\beta_{10}' + (1-\nu)^2\alpha_{20}\beta_{20}' + \nu(1-\nu)\Big(\alpha_{10}\beta_{20}' + \alpha_{20}\beta_{10}'\Big) \\ &= \begin{pmatrix} \nu^2\alpha_1\beta_1' & \nu(1-\nu)\alpha_1\beta_2' \\ \nu(1-\nu)\alpha_2\beta_1' & (1-\nu)^2\alpha_2\beta_2' \end{pmatrix},\end{aligned}$$

which means that there exists no ν that will result in linear combinations of α_{10}, α_{20} and β_{10}, β_{20}, respectively, which will provide a reasonable estimate of $\alpha\beta'$, since we cannot exclude the $\alpha_2\beta_1', \alpha_1\beta_2'$ products without eliminating either $\alpha_1\beta_1'$ or $\alpha_2\beta_2'$. Thus, in the case of two clusters with a misspecified rank, equal to half of the true rank, it is not possible to obtain a reasonable estimate of the block diagonal structure of Π. However, if we consider at higher dimensional model and try to reproduce the structure, we can allow for a rank estimate that is nearer, as a ratio, to the true value. Assume therefore, a system where

$$\alpha = \begin{pmatrix} \alpha_1 & 0 & 0 \\ 0 & \alpha_2 & 0 \\ 0 & 0 & \alpha_3 \end{pmatrix}$$
$$\beta = \begin{pmatrix} \beta_1 & 0 & 0 \\ 0 & \beta_2 & 0 \\ 0 & 0 & \beta_3 \end{pmatrix},$$

and define

$$\alpha_{10} = \begin{pmatrix} \alpha_1 & 0 \\ 0 & \gamma^{1/2}\alpha_2 \\ 0 & 0 \end{pmatrix} \qquad \alpha_{20} = \begin{pmatrix} 0 & 0 \\ (1-\gamma)^{1/2}\alpha_2 & 0 \\ 0 & \alpha_3 \end{pmatrix}$$
$$\beta_{10} = \begin{pmatrix} \beta_1 & 0 \\ 0 & \gamma^{1/2}\beta_2 \\ 0 & 0 \end{pmatrix} \qquad \beta_{20} = \begin{pmatrix} 0 & 0 \\ (1-\gamma)^{1/2}\beta_2 & 0 \\ 0 & \beta_3 \end{pmatrix},$$

where $\gamma \in [0, 1]$, such that

$$\begin{aligned} \alpha_{10}\beta_{10}' + \alpha_{20}\beta_{20}' &= \begin{pmatrix} \alpha_1\beta_1' & 0 & 0 \\ 0 & \gamma\alpha_2\beta_2' & 0 \\ 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & (1-\gamma)\alpha_2\beta_2' & 0 \\ 0 & 0 & \alpha_3\beta_3' \end{pmatrix} \\ &= \alpha\beta'. \end{aligned}$$

Defining as before, the estimates $\hat{\alpha}, \hat{\beta}$ as linear combinations $\hat{\alpha} = \nu\alpha_{10} + (1-\nu)\alpha_{20}$ and $\hat{\beta} = \nu\beta_{10} + (1-\nu)\beta_{20}$, respectively, we find that

$$\begin{aligned} \hat{\alpha}\hat{\beta}' &= \nu^2\alpha_{10}\beta_{10}' + (1-\nu)^2\alpha_{20}\beta_{20}' + \nu(1-\nu)(\alpha_{10}\beta_{20}' + \alpha_{20}\beta_{10}') \\ &= \begin{pmatrix} \nu^2\alpha_1\beta_1' & 0 & 0 \\ 0 & \big(\nu^2\gamma + (1-\nu)^2(1-\gamma)\big)\alpha_2\beta_2' & 0 \\ 0 & 0 & (1-\nu)^2\alpha_3\beta_3' \end{pmatrix} \\ &\quad + \nu(1-\nu)\begin{pmatrix} 0 & (1-\gamma)^{1/2}\alpha_1\beta_2' & 0 \\ (1-\gamma)^{1/2}\alpha_2\beta_1' & 0 & \gamma^{1/2}\alpha_2\beta_3' \\ 0 & \gamma^{1/2}\alpha_3\beta_2' & 0 \end{pmatrix} \end{aligned}$$

If $\nu = 0.5$, then $\hat{\alpha}\hat{\beta}' = \frac{1}{4}\alpha\beta' +$ "something", and since α, β are only determined up to a scaling constant, we see that it is possible to recover the structure of $\Pi = \alpha\beta'$, with the addition of some errors due to the lower dimensional approximation of α and β. However, note that, besides the block diagonal, the upper-right and lower-left corners of this error term are exactly zero, hence we find the estimated Π as

$$\hat{\Pi} = \begin{pmatrix} \hat{\Pi}_{11} & (1-\gamma)^{1/2}\hat{\Pi}_{12} & 0 \\ (1-\gamma)^{1/2}\hat{\Pi}_{21} & \hat{\Pi}_{22} & \gamma^{1/2}\hat{\Pi}_{23} \\ 0 & \gamma^{1/2}\hat{\Pi}_{32} & \hat{\Pi}_{33} \end{pmatrix},$$

where the magnitude of the entries of $\hat{\Pi}_{ij}$ for $i \neq j$ are scaled by $\gamma^{1/2}$ or $(1-\gamma)^{1/2}$ compared to $\hat{\Pi}_{ii}, i = 1, \ldots, 3$. Continuing this path of reasoning for examples with more than three clusters, it is evident, that due to the orthogonal structure between clusters, we can reasonably well estimate Π with $\hat{\Pi}$, when the estimation rank is lower than the true rank. The quality of the estimation is of course dependent on how close the estimation rank is to the true rank, the closer the better. This can be visualized by defining measure of "closeness" for two subspaces as defined by the matrices Π and $\hat{\Pi}$. For this purpose, one can use a generalized version of the vector angle, using the Frobenius inner product $\langle U, V \rangle_F = \mathrm{Tr}(U'V)$ for matrices $U, V \in \mathbb{R}^{p\times p}$, where $\mathrm{Tr}(\cdot)$ denote the trace operator. The angle between the matrices

U and V is then defined as

$$\Theta(U,V) = \arccos\left(\frac{\langle U,V\rangle_F}{\sqrt{\langle U,U\rangle_F\langle V,V\rangle_F}}\right). \tag{II.23}$$

For $U = 0$ and/or $V = 0$ we define $\Theta(U,V) = \pi/2$, i.e., the 0 matrix is always orthogonal to itself and any other. Figure II.2 displays $\Theta(\hat{\Pi},\Pi)$ for varying estimation rank $\hat{r}$, based on the same data as was used to produce Figure II.1. It is clear, that

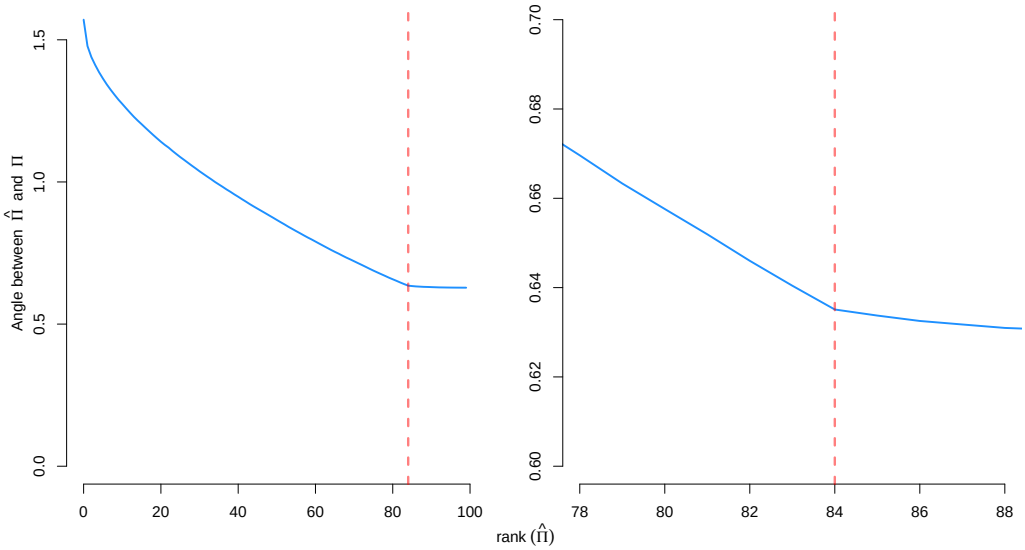

FIGURE II.2: Left: decreasing angle (blue) between $\hat{\Pi}$ and Π as the rank of $\hat{\Pi}$ approaches the true rank (red dashed line) of Π. When $\text{rank}(\Pi) < \text{rank}(\hat{\Pi})$, the angle appears close to an an asymptotic value. Right: same plot as the left, but focused on $r = 78,\ldots,88$.

as $\hat{r}$ increases, $\Theta(\hat{\Pi},\Pi)$ decreases from orthogonality to an asymptotic limit, which is reached at $\hat{r} = r$. Thus, overestimating $\hat{r}$, such that $r \leq \hat{r}$, implies that $\text{sp}(\Pi) \subseteq \text{sp}(\hat{\Pi})$ and in this case we cannot obtain a much better estimate, as measured by the matrix angle, of Π, than for $\hat{r} = r$. However, more importantly, Figure II.2 demonstrates that severely underestimation of the rank may have critical impact on reproducing the structure of Π, as described with the examples above, but in higher dimensions we can obtain fair estimates of Π, even if the rank is somewhat underestimated.

5 Simulation

Consider a system with $p = 100$ with twelve 8-clusters and four 1-clusters, i.e., a system of 12 coupled clusters each of size 8 and 4 independent processes. We let each cluster be coupled with strength $c_i = a_i b_i = 49^{-1}$ and assume that $\Omega = I_p$. A simulation of $N = 2000$ observations with initial condition $y_0 = 0$ is presented in Figure II.3, where the coloring is for easy identification of the 8-clusters (blue), the 1-clusters (red) and a pure 100-dimensional random walk for reference. It is noticeable how each cluster, as a whole, behave like a random walk. It is hard to tell from Figure II.3, but it might seem like there are only a few blue processes represented by thick lines. This is actually due to the overlapping of the individual processes intrinsic to each cluster. Here the coupling strength is strong enough that the internal cluster

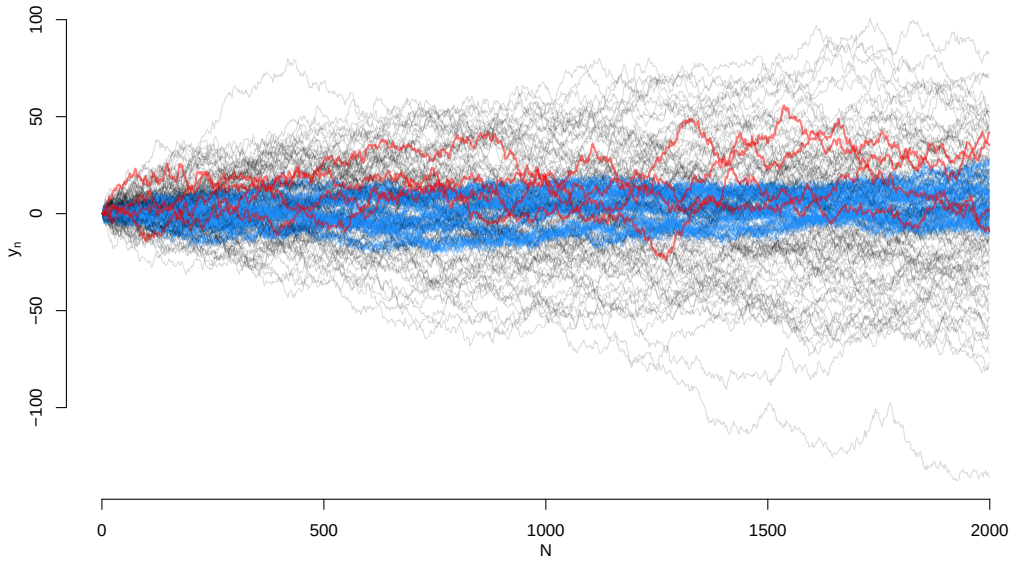

FIGURE II.3: Simulation of a linear Kuramoto system with 12 8-clusters (blue) and 4 1-clusters (red). A multivariate random walk (black) is superimposed for reference.

processes never stray far from the cluster, and as a whole they behave as a stationary process. The random walk reference shows how tightly the cointegrated processes are knitted.

5.1 Estimation of Π

The rank estimation of Π was performed with bootstrapping. From 300 bootstrap samples, the rank was determined as $\hat{r} = 82$. As such, the bootstrap test underestimates the rank, although by a small amount. Figure II.4 display the true Π matrix along with the maximum likelihood estimate $\hat{\Pi}$, under the assumption of $\hat{r} = 82 < r = 84$. Blue colors indicate positive values whereas red indicate negative values. As such, the diagonals are red and the blue squares represent the individual clusters. Note the lower right corner, where the 4 1-clusters are represented as 0 entries in Π.

As discussed in Section 4, for high-dimensional systems it is of less importance to hit the true rank exactly, compared to low-dimensional systems. The estimate $\hat{\Pi}$ in Figure II.4, given $\hat{r} = 82 < r = 84$ visually reproduces much of the structure of Π, and as was shown in Figure II.2, the angle between Π and $\hat{\Pi}$ is only slightly reduced for $\hat{r} = 84$ versus $\hat{r} = 82$. However, it is somewhat surprising, that $\hat{\Pi}$ visually captures much the true structure even if the rank is severely under-/overestimated. Figure II.5 shows $\hat{\Pi}$ given $\hat{r} = 74 = r - 10$, $\hat{r} = 84 = r$ and $\hat{r} = 94 = r + 10$, i.e., for estimation rank at the true value $r = 84$ against $\mathrm{rank}(\hat{\Pi}) = r \pm 10$. In all three cases, the cluster structure of Π is reproduced quite well with noticeable distinct 8-clusters and 1-clusters.

5.2 Estimation under structural restrictions

We compare 4 different estimations of Π for the 2000 simulated observations of (II.1) (see Section 5). All estimators are under the assumption of $H_r : \mathrm{rank}(\Pi) = r = 84$,

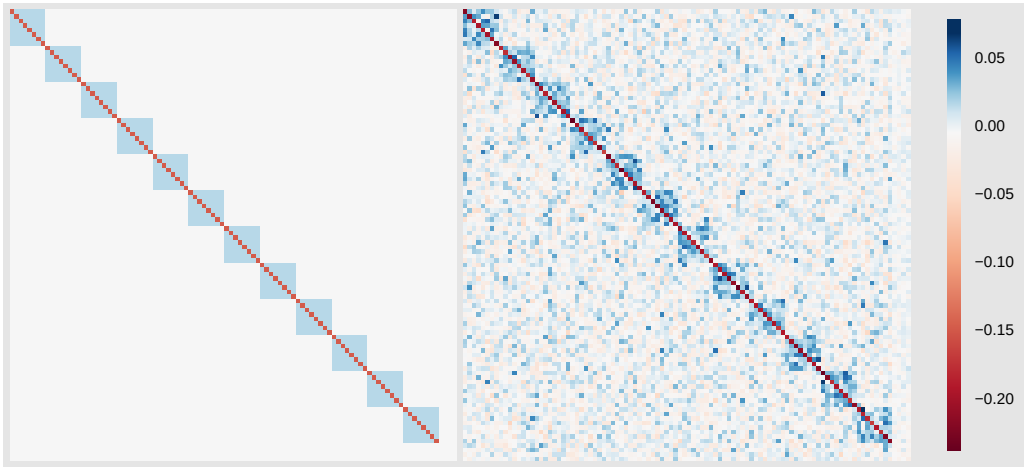

FIGURE II.4: Estimating Π with $\hat{r} = 82$, given that the true rank is $r_{\text{true}} = 84$.

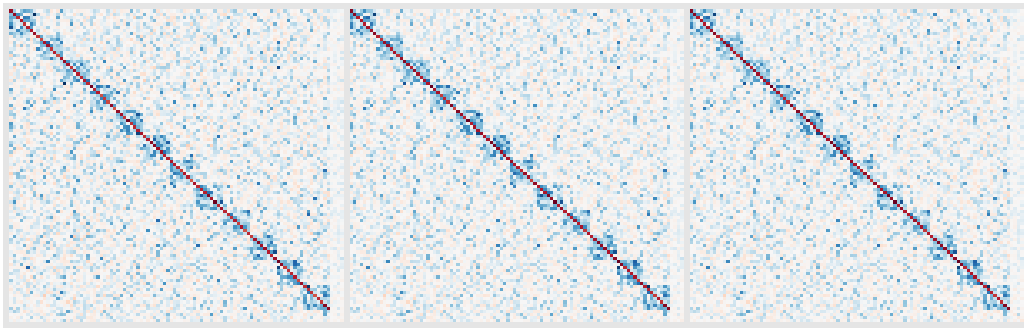

FIGURE II.5: Estimation of Π. Left: $\hat{\Pi}$ for $\hat{r} = 74 = r_{\text{true}} - 10$. Middle: $\hat{\Pi}$ for $\hat{r} = 84 = r_{\text{true}}$. Right: $\hat{\Pi}$ for $\hat{r} = 94 = r_{\text{true}} + 10$

i.e., we assume that the rank is known and equals the true rank of the simulated system.

The first estimator is the unrestricted maximum likelihood estimate $\hat{\Pi} = \hat{\alpha}\hat{\beta}'$, as derived by Johansen (1996). Note that this is not restricted to be symmetric. The second estimator will be the penalized OLS estimator

$$\hat{\Pi}_{\text{pen}} : \min_{\Pi} \left|\left| S_{01}S_{11}^{-1} - \Pi + \kappa \sum_{ij} |\Pi_{ij} - \Pi_{ji}| \right|\right|_F,$$

where $\kappa = 20$. The third estimator is the projection of $\hat{\Pi} = \hat{\alpha}\hat{\beta}'$ onto S_p

$$\hat{\Pi}_{\text{proj}} = \frac{1}{2}(\hat{\alpha}\hat{\beta}' + \hat{\beta}\hat{\alpha}'),$$

and the final estimator is the symmetric OLS

$$\hat{\Pi}_{\text{sym}} = \frac{1}{2}(S_{01}S_{11}^{-1} + S_{11}^{-1}S_{10}).$$

The estimators $\hat{\Pi}_{\text{pen}}$, $\hat{\Pi}_{\text{proj}}$ and $\hat{\Pi}_{\text{sym}}$ are transformed to their corresponding low rank approximations using (II.7) and $r = 84$. Note that $\hat{\Pi}_{\text{proj}}$ is similar to he lift-and-project algorithm.

We compare the estimates with four measures: i) symmetry, by $||M - M'||_F$, ii) the matrix angle (II.23) between the true Π and a given estimator, iii) the norm distance $||\Pi - M||_F$ and iv) the standard deviation of the entries outside the block diagonal of the true Π. The last measure gauges the size of the entries in the estimates that are precisely zero in the true Π.

Estimator	Symmetry	Angle	Distance	Std.Dev	$-2\log Q(\hat{\Pi}_{\text{est}}, \hat{\Pi})$
$\hat{\Pi}$	1.9663	0.6352	1.5631	0.0147	
$\hat{\Pi}_{\text{pen}}$	0.0083	0.5020	1.2251	0.0109	16018.96
$\hat{\Pi}_{\text{proj}}$	0	0.4966	1.2049	0.0107	14215.74
$\hat{\Pi}_{\text{sym}}$	0	0.4927	1.2009	0.0107	13618.04

TABLE II.1: Comparison of various estimators of Π, based on $N = 2000$ observations from a $p = 100$ dimensional system.

It is evident from Table II.1 that $\hat{\Pi}_{\text{proj}}$ and $\hat{\Pi}_{\text{sym}}$ are the best choices among the estimators here, with $\hat{\Pi}_{\text{sym}}$ slightly superior to $\hat{\Pi}_{\text{proj}}$. The angle between these two estimators is $\Theta(\hat{\Pi}_{\text{proj}}, \hat{\Pi}_{\text{sym}}) = 0.0251$ and the norm distance is $||\hat{\Pi}_{\text{proj}} - \hat{\Pi}_{\text{sym}}||_F = 0.0575$, hence the estimators are near each other in the context of the measures considered here. However, considering the likelihood ratio test statistics ($-2\log Q(\hat{\Pi}_{\text{est}}, \hat{\Pi})$), the $\hat{\Pi}_{\text{sym}}$ has a lower test statistic compared to both of the others. Figure II.6 presents a histogram based on 1000 simulations of the difference of the $-2\log Q$ statistics $-2\log Q(\hat{\Pi}_{\text{proj}}, \hat{\Pi}) - \big(-2\log Q(\hat{\Pi}_{\text{sym}}, \hat{\Pi})\big)$, from which it is evident that the $\hat{\Pi}_{\text{sym}}$ esti-

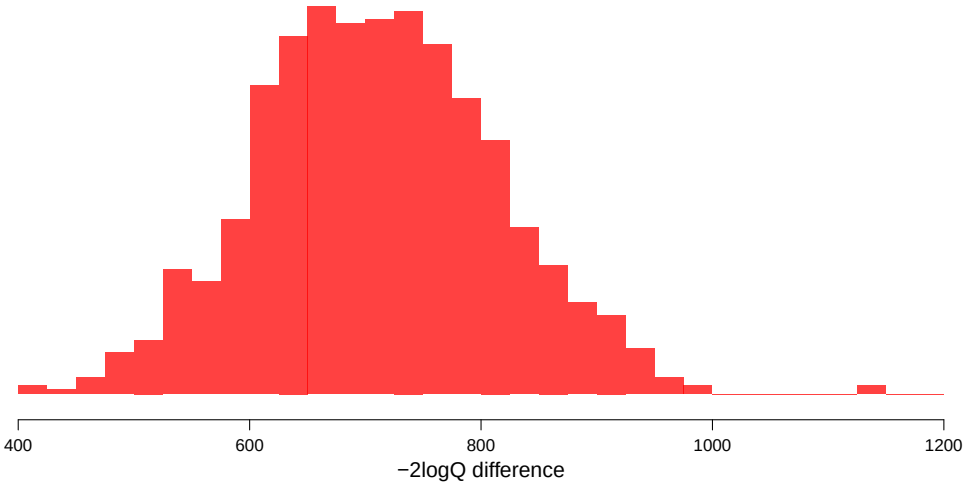

FIGURE II.6: Histogram.

mator has a lower likelihood ratio test statistic, compared to the $\hat{\Pi}_{\text{proj}}$ estimator. This implies that the symmetric projection of the OLS estimator is the optimal choice among the three estimators considered here, when compared to the unrestricted $\hat{\Pi}$ estimator.

Figure II.7 displays the true Π matrix versus the unrestricted maximum likelihood estimator $\hat{\Pi}$ (top right) and the two symmetric estimators $\hat{\Pi}_{\text{proj}}$ (bottom left) and $\hat{\Pi}_{\text{sym}}$ (bottom right). By eyeballing Figure II.7, it is evident that the symmetric estimators exhibit much less noise in the values away from the block diagonal, which is also clear from the standard deviation measure in Table II.1. The structure of Π is

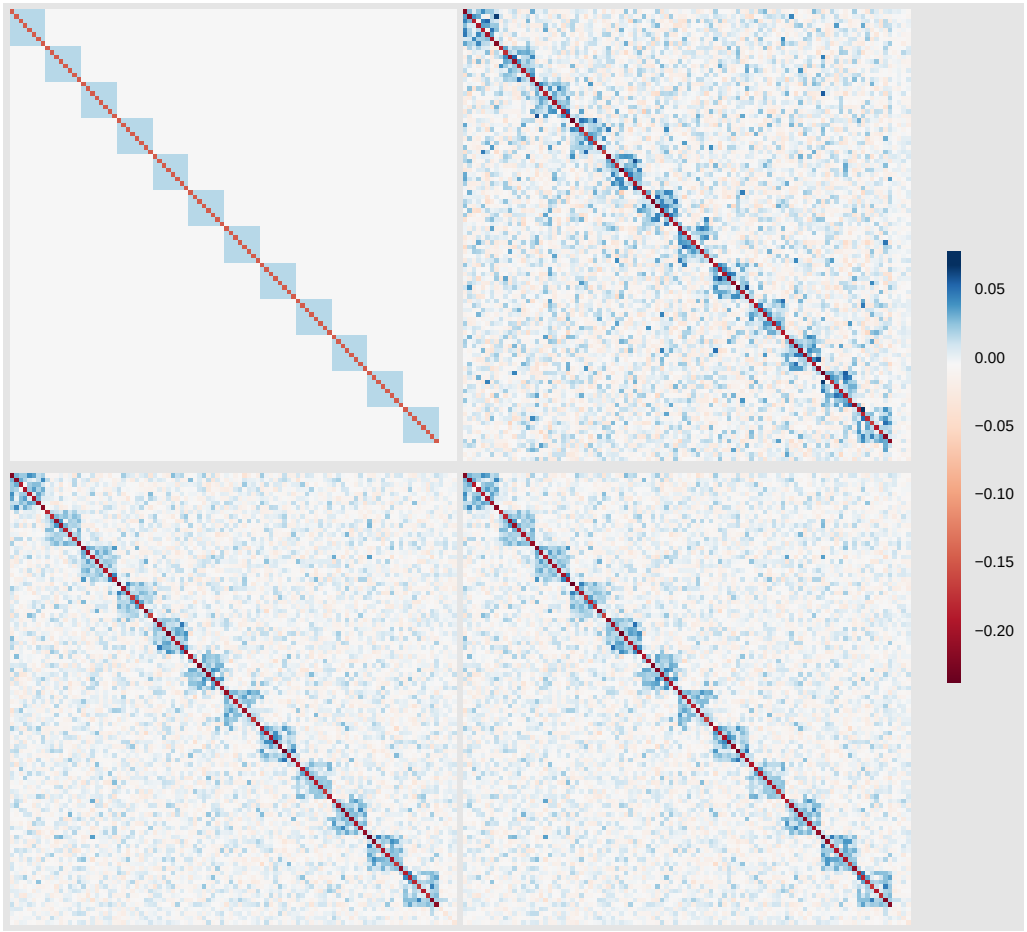

FIGURE II.7: True, unrestricted and symmetric Π estimates under $H_r : \mathrm{rank}(\Pi) = r = 84$. Topleft: True Π matrix. Topright: unrestricted $\hat{\Pi}$. Bottomleft: $\hat{\Pi}_{\text{proj}}$. Bottomright: $\hat{\Pi}_{\text{sym}}$. The bottom estimators look very similar.

much more clearly defined in the symmetric estimators, thus a likelihood ratio test for the linear Kuramoto type system with Π as in (II.17) should utilize these for the restriction of the model to $H_r^s \subset H_r$, i.e., symmetry under H_r.

The system considered here relies heavily on the fact, that the observed processes have been ordered according to the cluster structure of true Π matrix. This might not be so in actual observations, hence it is necessary to perform some form of clustering before testing any symmetry in the system. Fortunately, having determined the rank r of the system, then $p - r$ is precisely the number of clusters (including 1-clusters) and thus running a cluster algorithm can be done with this input.

6 Discussion

With this manuscript we have examined cointegration in a high-dimensional setting. By defining a linear cluster version of the Kuramoto (1984) model, we have assessed the capability of the bootstrap determination of the rank in such a system as well as new estimators in restrictions to a symmetric cointegration matrix. The

bootstrap procedure was found to be efficient in approximating the rank in a high-dimensional setting, although it is prone to underestimate the rank due to the standard procedure of starting with the hypothesis of no cointegration ($r = 0$) and then going through hypotheses where the rank is increased until the hypothesis cannot be rejected. However, it was demonstrated that as the dimension increased, determining the precise rank became less important than in low-dimensional settings. This was due to the fact that in high-dimensional settings, determining the rank up to some relative proportion of the true rank admitted a representation in a lower (than the true model) dimensional space that approximated the original structure quite well (Figure II.2), when measured in terms of the matrix angle (II.23). Using a rank lower than the true rank, the usual reduced rank estimator was found to visually reproduce most of the true structure from the simulated model. In extension to this we also demonstrated various symmetric estimators for the system, which produced superior results to the unrestricted reduced rank estimator. From theoretical considerations and numerical results, the symmetric projection of the standard OLS estimator, approximated with a low rank matrix using singular value decomposition, was found to be the optimal choice among the estimators considered here, according to the measures used in Table II.1 and it is our conjecture that this is indeed the best estimator, given a symmetric low rank system, with respect to these measures. One of the deficiencies highlighted in the results, was the fact that the simulated data were ordered correctly when used as input for the estimation. However, in a real data context the order is unknown. As such, the work presented here should be extended with efficient methods to first order data, such that the symmetric estimator may be applied. As mentioned above, one way is to perform a clustering of the processes, prior to symmetric estimation. Due to the fact that the rank can be assumed known before clustering, the number of clusters are already known. This would be an important prior knowledge in order to pick an efficient clustering algorithm. The full analysis of a symmetric system would then be as follows: i) determine the rank, ii) cluster the data and iii) test for symmetry. Other interesting extensions to this work is the unused fact that given a certain cluster structure, the remaining entries of an estimator should be zeros. Hence, there is a large amount of sparsity in the linear systems considered here. Therefore, including sparsity constraints would be of interest to further improve the estimation procedure. Finally, testing for the complete indepence among clusters in a high-dimensional system would also be of interest, since this would basically allow the problem to be split into a number of lower dimensional settings. This would require indepence among the stochastic trends in the system, or in other words sparsity constraints for the full covariance matrix.

Spike train modeling

CHAPTER 3

Spike train modeling

1 Introduction

This part will focus on the analysis of binary spike train data using Generalized Linear Models (GLMs). The contributions in this part consist of two manuscripts

- Capturing spike variability in noisy Izhikevich neurons using point process Generalized Linear Models. (Østergaard, Kramer, and Eden, 2018)
- A GLM State Space Model for Bursting Neurons. (Østergaard et al., 2017)

The first paper, accepted for publication in Neural Computation (Østergaard, Kramer, and Eden, 2018), describes how well GLMs capture intrinsic information as well as variability within observed spike trains, for simulated Izhikevich neurons of various types, injected with noisy currents. The research idea was to pick a wellknown generative neuron model that was easy to implement and could produce various complex behaviors observed in both *in vitro* and *in vivo* recordings of neurons. This model was not initially constructed with a stochastic component, but by augmenting it with a noise component through the injected current, we were able to control the amount of randomness in the model. Then, for each type of neuron and noise setting, we investigated the fitted statistical GLM in terms of a goodness-of-fit analysis, to assess how well the GLM actually captured both the neuron-specific properties, such as refractoriness, as well as variability in the spike trains. We designed a history dependent multiplicatively separable GLM, using indicator basis functions, to include the history of the observed spike train as explanatory variables. We found this model to be very flexible in terms of describing the various types of neurons as well as the varying noise, but in the deterministic or near deterministic setting, the random component of the GLM is not captured properly, thus resulting in a poor model fit in a statistical sense. Hence, the model might still be able to capture intrinsic information in the spike train, but the generalization properties would be quite restricted. In addition, we found that capturing bursting was, not surprisingly, more complex than capturing tonic spiking, i.e., spike trains with a unimodal InterSpike-Interval (ISI) distribution.

The second manuscript is a working paper, to be submitted shortly. This paper deals explicitly with modeling the bimodal ISI distributions of bursting neurons. Here the idea is to augment the simple history dependent and multiplicatively separable GLM from Østergaard, Kramer, and Eden (2018) with a latent state to control

for the dual behavior of bursting and resting (in between bursts) respectively. The model's capability of explicitly modeling the dual behavior enables it to capture the two differing refractory periods between spikes when resting and when bursting, producing the bimodal ISI distribution. On a more abstract level, there is nothing which inhibits a different interpretation of the latent state and thus the modeling of ISI distributions with more than two modes. As such, the model can also easily be extended to include other features, such as adaptive spiking behavior. The paper demonstrates how a multiplicatively separable and history dependent GLM model can be enhanced with a latent state and how to apply a marginalized particle filter for point process observations to simultaneously estimate the GLM parameters and decode the latent state of bursting. The results in the paper display the models abilities for simulated Izhikevich neurons with injected noise as in Østergaard, Kramer, and Eden (2018). Furthermore a preliminary connection to the Izhikevich phase space is established by showing how the decoded bursting state identifies when the Izhikevich variables crosses the separatrix of stable/unstable regions.

The first paper was written in collaboration with Uri Eden and Mark Kramer, as part of a research stay in Boston from November 2016 to May 2017. The second paper is an extension of this work in collaboration with Uri Eden and Susanne Ditlevsen.

2 GLMs for spike train data

This overview of GLMs and point processes is based on Truccolo et al. (2005) and Kass, Eden, and Brown (2014).

The Generalized Linear Model (GLM) is a fundamental part of any statisticians toolbox. This model framework consist of a broad class of models which are very flexible and can be adapted to many different settings. Here we review GLMs particularly suited for binary spike train analysis, leading to the class of history dependent, multiplicatively separable GLMs with a Poisson random component.

The GLM describes the relation between an observation $Y \in \mathbb{R}$ and explanatory variables $x \in \mathbb{R}^p$ as

$$g(\mathbb{E}_{\mathcal{P}}[y]) = x\beta, \tag{3.1}$$

with three main components

1. link function $g(\cdot)$,
2. systematic component: the linear predictor $x\beta$,
3. random component: a relevant distribution $\mathcal{P}$ from the *exponential family*.

In Truccolo et al. (2005) the authors derive a GLM framework for point processes with applications to spike train data. Particularly, the authors show that a Binomial or Poisson distribution can be used for this task.

Assuming that we observe a neuron in some interval Δt, we can interpret the neurons *firing rate* as

$$FR = \frac{\text{\# of observed spikes}}{\Delta t}. \tag{3.2}$$

However, (3.2) is a vague interpretation of the firing rate, so instead we could ask: *what is the expected number of spikes in the next observation period*? A spike train is a 0-1 valued time series, so if we let $\Delta t \to 0$ we either observe 0 or 1 for Δt small enough. Hence the expected number of spikes is equal to the probability parameter of a Bernoulli variable and can be formalized as $P(\text{spike in } t, t + \Delta t)$. Scaling by Δt such that the interpretation is not dependent on the size of Δt, and letting $\Delta t \to 0$, the *instantaneous firing rate* is obtained

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(\text{spike in } t, t + \Delta t)}{\Delta t}. \tag{3.3}$$

The probability in (3.3) may be time dependent, but it does not account for any other factors and hence it is of little practical use. Hence, conditioning on other factors, denoted here by x_t, and a set of parameters θ, we get the *conditional firing rate*

$$\lambda(t|x_t, \theta) = \lim_{\Delta t \to 0} \frac{P(\text{spike in } t, t + \Delta t|x_t, \theta)}{\Delta t}. \tag{3.4}$$

2.1 Point process Representations

In order to model the probability in the numerator of (3.4), we turn to the theory of *point processes* which is the class of stochastic processes that deal with events (points) in continuous time t. Biological processes, such as membrane potentials, are rightly interpreted as continuous time processes. We can therefore interpret spike trains, which is a set of spike times for a neuron $(s_1, \ldots, s_n)$, where $s_1 < s_2 < \cdots < s_n$, as the jumptimes for a continuous time process. Notice that with the strict inequalities we implicitly assume that we observe at most one spike at any given time. If we assume that these observations, or events, originate from the underlying process $N(t)$, which we here define as the number of spikes counted up to time t. Hence, $N(t)$ is a counting process with increments at each spike time $s_k, k = 1, \ldots, n$. The process $N(t)$ called *orderly* if

$$\lim_{\Delta t \to 0} \frac{P(N(t + \Delta t) - N(t) > 1|x_t, \theta)}{\Delta t} = 0,$$

implying that for sufficiently small Δt, the probability of observing more than one spike in the interval $(t, t + \Delta]$ is negligible. Another continuous representation of the spike times is the waiting time between two spike times $w_k = s_k - s_{k-1}$. These waiting times are also termed Inter-Spike Intervals (ISI), and the distribution of these will often clearly reveal intrinsic neuron features such as refractoriness. However, since it is practically impossible to observe any process continuously, spike trains are instead recorded at some frequency of observation times which implies that the time variable is binned in equidistant intervals of size $\Delta t = t_k - t_{k-1}$. Assuming that the time resolution Δt is sufficiently small, such that we observe at most one spike in any timebin, then we can represent the spike train as a binary time series with observations at times $t_1 < t_2 < \cdots < t_n$. The representations mentioned above are visualized in Figure 3.1 which is reproduced from Kass, Eden, and Brown (2014). The binary time series representation is modeled by the discretized process $\Delta N_k = N(t_k) - N(t_{k-1})$ and is often the form of recordings with $\Delta t = 1$ ms or similarly in the order of ms. Denoting $Y_k = \Delta N_k$, then within each time bin, Y_k follows a Bernoulli distribution with some corresponding probability parameter p_k. Thus, the binary time series $Y_k, k = 1, \ldots, n$ can be viewed as a sequence of Bernoulli

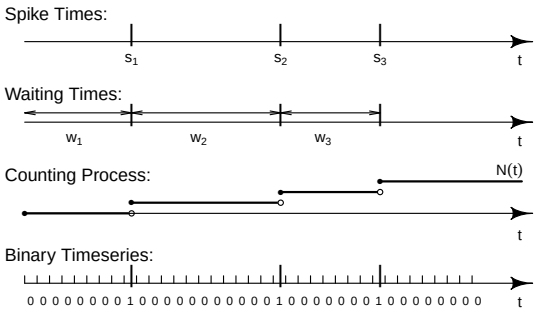

FIGURE 3.1: Various representations of spike train data. Reproduced from Kass, Eden, and Brown (2014).

random variables where p_k can vary across time bins. In this case the point process is called *inhomogeneous*, whereas if p_k is constant across time bins, the point process is *homogeneous*.

2.2 Poisson GLM for spike trains

Given an orderly counting process $N(t), t \in (0, T]$, approximated by the discrete sequence of Bernoulli variables, $Y_k, k = 1, \ldots, K$, we define the likelihood function of K i.i.d. observations as the joint probability mass function of the observations $\{y_k\}_{k=1}^K$

$$p(y_1, \ldots, y_K) = \prod_{k=1}^{K} p_k^{y_k} (1 - p_k)^{1-y_k}, \tag{3.5}$$

where each probability p_k can be interpreted in terms of the conditional instantaneous firing rate (3.4) for Δt small

$$p_k \approx \lambda(k|\, x_k, \theta)\Delta t, \tag{3.6}$$

where $\lambda(k|\, x_k, \theta) = \lambda(t_k|x_{t_k}, \theta)$. Also, when Δt is small then $1 - p_k \approx \exp(-p_k)$ and $\log(p_k(1 - p_k)^{-1}) \approx \log(p_k)$ such that

$$\begin{aligned}
\prod_{k=1}^{K} p_k^{y_k} (1 - p_k)^{1-y_k} &= \prod_{k=1}^{K} \Big(\frac{p_k}{1 - p_k}\Big)^{y_k} (1 - p_k) \\
&\approx \prod_{k=1}^{K} \exp\left[y_k \log\Big(\frac{p_k}{1 - p_k}\Big) \right] \exp(-p_k) \\
&= \exp\left[\sum_{k=1}^{K} \left(y_k \log\Big(\frac{p_k}{1 - p_k}\Big) - p_k \right) \right] \\
&\approx \exp\left[\sum_{k=1}^{K} \Big(y_k \log(p_k) - p_k \Big) \right].
\end{aligned}$$

Assuming J observed spikes in $(0, T]$ and using the relation (3.6), this leads to a discrete approximation of the continuous time likelihood for the Poisson process

with (continuous) intensity function $\lambda(t|x_t, \theta)$ (see Truccolo et al., 2005)

$$\begin{aligned}\frac{1}{(\Delta t)^J}\exp\left[\sum_{k=1}^{K}\Big(y_k\log(p_k)-p_k\Big)\right] &\approx \frac{1}{(\Delta t)^J}\exp\left[\sum_{k=1}^{K}\Big(y_k\log\big(\lambda(k|x_k,\theta)\Delta t\big)-\lambda(k|x_k,\theta)\Delta t\Big)\right]\\ &= \exp\left[\sum_{k=1}^{K}\Big(y_k\log\lambda(k|x_k,\theta)-\lambda(k|x_k,\theta)\Delta t\Big)\right]\\ &\underset{\Delta t\to 0}{\longrightarrow}\exp\left[\int_0^T\log\lambda(t|\,x_t,\theta)dN(t)-\int_0^T\lambda(t|x_t,\theta)dt\right].\end{aligned}$$

Note that in the above we exploit that $\sum_{k=1}^{K} y_k \log\lambda(k|x_k,\theta)\Delta t = \sum_{k=1}^{J} y_k \log\lambda(k|x_k,\theta)\Delta t$, due to the defintion of y_k. Hence, we obtain a discrete approximation of the log-likelihood function

$$L(y_1,\ldots,y_K)=\sum_{k=1}^{K}\Big(y_k\log\lambda(k|x_k,\theta)-\lambda(k|x_k,\theta)\Delta t\Big), \tag{3.7}$$

for the binary time series observations $\{y_k\}_{k=1}^{K}$.

For members of the exponential family the distribution function have the general form

$$f(y,\theta,\phi)=\exp\left(\frac{\theta y-b(\theta)}{a(\phi)}+c(y,\phi)\right), \tag{3.8}$$

with canonical parameter θ, dispersion parameter ϕ and $a(\cdot), b(\cdot)$ are functions determined by the specific distribution class. For a Poisson distribution with parameter λ, the canonical parameter is $\log\lambda$, $a(\phi)=1$, $b(\theta)=\exp(\theta)=\lambda$ and $c(y,\phi)=-\log(y!)$, sucht that (3.8) takes the well known form

$$f(y,\theta,\phi)=\exp\left(\frac{y\theta-b(\theta)}{a(\phi)}+c(y,\phi)\right)=\lambda^y\frac{e^{-\lambda}}{y!}$$

This reveals that the canonical link for the Poisson GLM is the $\log$ function, and hence that interpreting the linear predictors $x\beta$ in (3.1) under the multiplicatively separable assumption, means that for each parameter $\beta_j, j=1,\ldots,p$, then $\exp(\beta_j)$ denotes the multiplicative adjustment of λ, or for some given baseline β_0, then $\exp(\beta_j)$ is the modulation of the baseline firing rate $\exp\beta_0$, given the corresponding predictor in x. This interpretation seem much more intuitive than using the Bernoulli interpretation. In this case, the components in (3.8) are $\theta=\log\frac{p}{1-p}$, $b(\theta)=\log(1+\exp(\theta))$, $a(\phi)=c(y,\phi)=1$ such that

$$f(y,\theta,\phi)=p^y(1-p)^{1-y}.$$

and the canonical link is then the logit function

$$\mathrm{logit}(p)=\log\left(\frac{p}{1-p}\right).$$

As such the interpretation of the linear predictors becomes in terms of log-odds, rather than multiplicative adjustments. Fortunately, we can rely on either model due to the convergence of the Bernoulli and Poisson distributions when $\Delta t\to 0$ as

explained above.

2.3 History dependent GLM

By choosing a relevant set of explanatory variables in (3.1), we can design a GLM to capture relevant features in spike train data. One such feature is refractoriness of the neuron, i.e., the suppression of any spike activity in the membrane potential following immediately after an observed spike. This implies that it is necessary to include information on when the last spike was observed or simply the trailing history of the spike train back to and including the last spike. If we write up the log of the conditional intensity function as

$$\log \lambda(k|\{y_j\}_{j=k-1}^{k-p}, \theta) = \beta_0 + \sum_{j=1}^{p} \beta_j y_{k-j}, \tag{3.9}$$

we include the trailing history of $y_{k-1}, y_{k-2}, \ldots, y_{k-p}$ as predictors for the intensity of y_k with the corresponding parameter $\theta = (\beta_0, \beta_1, \ldots, \beta_p)' \in \mathbb{R}^{p+1}$. From (3.9) it is evident that since $y_{k-j} \in \{0, 1\}$, each $\beta_j, j = 1, \ldots, p$ correspond to the modulation of the baseline firing rate $\exp(\beta_0)$ as mentioned above, given the past spiking history. Thus, for a neuron, the refractoriness should effectively amount to low β_j estimates, for low values of j, since this will suppress the firing rate following a spike ($y_k = 1$), and gradually as the previous spike is further and further away, the β_j's should increase to indicate an increased firing probability. It is of course straight forward to include other factors in (3.9), such as stimulus effects and other spatial, temporal or ensemble effects, revealing the flexibility of the GLM framework to capture multiple features of spike trains.

2.4 Basis functions for GLM regression

An unfortunate consequence of (3.9) is, that in order to capture any historical dependencies, such as refractoriness, it is necessary to include a fair amount of history. This leads to a considerable amount of parameters included in the model, many of which are possibly redundant, or at the very least highly correlated. Hence, in order to obtain a tractable model, it is necessary to reduce the dimension, while keeping the amount of trailing history fixed. Assuming that we want to include the last m observations $(y_1, \ldots, y_m)'$ as explanatory variables in (3.9), we can reduce the dimensionality of the model by utilizing basis functions to model the structure of the history. More formally, if we let $B \in \mathbb{R}^{m \times p}$ denote a linear transformation from an m-dimensional space into a p-dimensional space, then we can rewrite (3.9) as

$$\log \lambda(k|\{y_j\}_{j=k-1}^{k-m}, \theta) = \beta_0 + \sum_{j=1}^{p} \beta_j \sum_{i=1}^{m} B_{ij} y_{k-i}, \tag{3.10}$$

where B_{ij} denote the i'th row and j'th column of B. Given the representation in (3.10), the model dimension becomes $p+1$ at the expense of underlying assumptions for the basis functions. Trivially, for $B = I_m$, then (3.9) and (3.10) coincide, thus choosing a proper basis amounts to choosing the right transformation B such that the dimension is reduced, while keeping assumptions realistic. The intuition behind basis functions is, that if we assume the trailing history of the spike train y_k has a

functional form $f(\tau), \tau \leq t_k$, we may represent f in terms of the basis B, much like a basis in a vector space

$$f(\tau) = \sum_{j=1}^{p} \beta_j B_{\cdot j}(\tau), \tag{3.11}$$

with $B_{\cdot j}$ denoting the $j'th$ column of B and β_j act as the weight for basis function j. Thus, using a basis in (3.10) implicitly assumes a functional form such as (3.11) of the past information in y_k.

A popular choice of basis functions for (3.11) are *spline functions*, which in extension to the functional form also assumes a continuous structure. Given the continuous nature of the underlying physiological process generating a spike train, this assumption does not seem unreasonable. Also, given the possibility of significantly reducing the number of parameters, this assumption is usually highly rewarding. The construction of the matrix B in terms of spline basis functions can be done from various choices of splines. General spline functions use a set of polynomials defined such that the tangents at the endpoints match to obtain a continuous curve composed of combinations of these basis functions. For *canonical splines*, these polynomials are cubic and the tangents at the endpoints are calculated with a tension parameter s

$$(1 - s)\frac{y_{k+1} - y_{k-1}}{t_{k+1} - t_{k-1}}, \tag{3.12}$$

where $s \in [0, 1]$. Note that if $s = 1$, then (3.12) is 0 and thus the connecting endpoints for the spline basis functions are restricted to horizontal tangents. When $s = 0$, (3.12) corresponds to the tangents for *Catmul-Rom* splines.

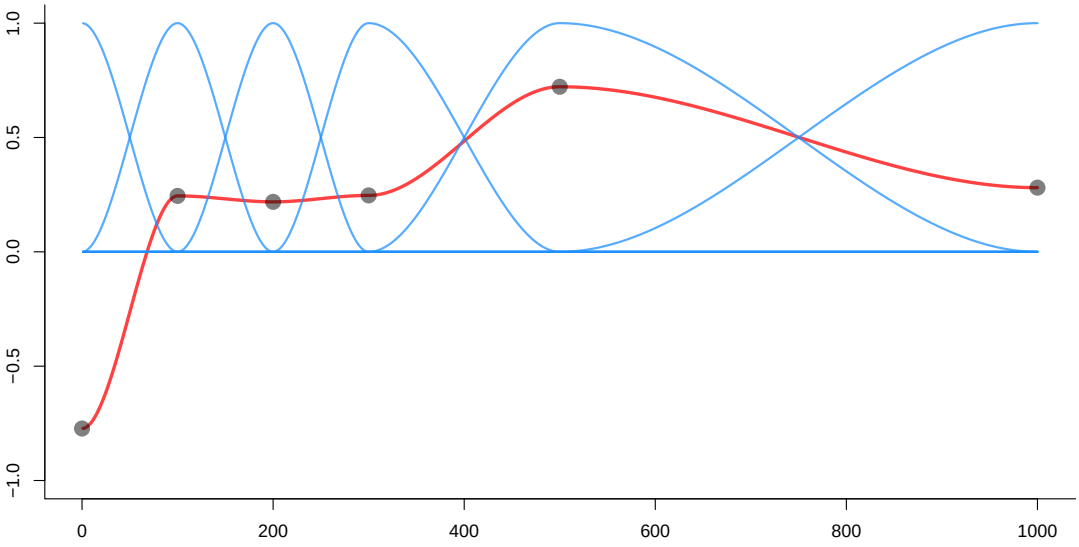

FIGURE 3.2: Example of spline basis functions and a random curve. The 6 basis functions (blue curves) make up the columns of the matrix B of size 1000×6. Given 6 weights (gray dots), these basis functions will construct a continuous curve (red).

To use splines in a regression context we must choose a set of *control points* $\mathcal{C} = (c_1, \ldots, c_r)$. These can be chosen based on an ISI histogram for a given spike train, but other means of choosing the set of points exist. We will, however, not pursue

these techniques here. To ensure a low dimensional representation of the trailing history, the number of basis functions should be kept to a minimum as there will be one parameter (weight) per basis function. Hence, by choosing control points around the areas of the ISI histogram where there is structure of interest, we capture the essential features of the neuron with respect to these areas of the trailing history. Figure 3.2 present a toy example of unequally spaced spline basis functions (blue) on the interval $[0, 1000]$ with control points (gray) $\mathcal{C} = (0, 100, 200, 300, 500, 1000)$ and $\bar{\beta} = (-0.77, 0.25, 0.22, 0.25, 0.72, 0.28)$ as arbitrary weights to construct a curve (red). The flexible positioning of control points and the number of basis curve to use provide a simple framework to capture features such as refractoriness of spike trains.

Compared to splines, a simpler choice of indicator basis functions can be used to represent the trailing history of a spike train. With these functions B becomes a binary matrix of 1-0 entries

$$B_{ij} = \mathbf{1}_{\left[(j-1)w+1, jw\right]}(i) = \begin{cases} 1 & \text{for } i = (j-1)w+1, \ldots, jw \\ 0 & \text{else} \end{cases}, \tag{3.13}$$

for some predetermined width w of each indicator. Plugging this into (3.10) we observe that each column j represent the total spike activity in each timebin which the basis covers $[(j-1)w, jw]$. As an example, if $\Delta t = 0.1$ ms and $w = 10$, then each indicator function covers 1 ms. The full matrix B will the simple structure

$$B = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix}.$$

where each column represent the cumulative spike activity in the corresponding time bin.

Using indicator basis functions is a simple choice that only requires setting the width w as opposed to the set of control points $\mathcal{C}$ for spline basis functions. While the reduction of the parameter dimension is not as effective as for splines, it allows for a more flexible model to represent the history than splines as it is closer to choosing the full GLM form (3.9) without any basis, while still reducing the number of parameters to some extend by a factor equal to w. However this also introduces the statistical dilemma regarding fewer/more parameters and estimation.

For spike trains recorded at high frequencies, this implies that multiple bins will most likely be zero and hence the time series will be very sparse. In this case, choosing an indicator basis might be more adequate than splines, if the interspike interval distribution resembles a delta function. It will reduce the need for individual parameters at each time bin at the cost of precision regarding when the past spikes actually

occurred. One can only say which bin, but not where in the bin the spike occurred. The use of indicator basis functions contrary to spline basis functions are also discussed in Østergaard, Kramer, and Eden (2018) with respect to approximating delta functions.

2.5 LASSO estimation

LASSO regression (Tibshirani, 1996) have become a popular choice of penalized regression to balance the number of parameters against model fit. Any given log-likelihood function $L(\theta)$ can be penalized by subtracting a term

$$L_\kappa(\theta) = L(\theta) - \kappa \sum_{j=0}^{p} ||\beta_j||^q, \tag{3.14}$$

such that $L(\theta)$ is restricted by the magnitude of parameters β_j. For $q = 1$ this is LASSO regression or L^1 penalization. Another popular choice is $q = 2$ which corresponds to an L^2 penalization termed ridge regression. The main difference between these methods is the emphasis on sparsity. Whereas an L^2 penalty will shrink parameters towards 0, the L^1 penalty will shrink (some of) them to precisely 0, effectively reducing the model dimension.

Besides imposing sparsity (when using L^1 penalization) another reason to introduce penalization of the parameters is to keep the magnitude of parameters under control. Contrary to an unrestricted likelihood optimization where the parameters may diverge or grow out of proportions such that numerical routines collapse, penalized likelihood optimization will ensure that this is kept under control. When some regions of a likelihood surface are almost flat in one or more directions, the reason why optimization algorithms fail is, that increasing a single parameter may continuously increase the likelihood. Although the marginal increase is insignificant, this will still cause a numerical optimization to break and return errors since it does not converge to an optimum. Introducing any penalization such as (3.14) will help to avoid this problem regardless of the choice of q.

Besides choosing q, the choice of the penalization parameter κ is crucial for this type of regression. The value of κ can be chosen by cross validation or by other means. In Østergaard, Kramer, and Eden (2018) the value is chosen with respect to the Kolmogorov Smirnov statistic (see Section 2.5) to emphasize model performance in a goodness-of-fit setting. Hence, the choice of κ can be modified with a certain perspective in mind, although cross validation is most commonly used.

3 State Space Models

State space models, also known as hidden Markov models, refer to a class of models, where the underlying dynamics (continuous or discrete) are controlled by a latent process x_t. While presenting the relevant theory for this thesis, we will assume that time is discrete (or discretized). Thus, the subscript k refers to the k'th observation at time t_k, $t_{k-1} < t_k$, and we refer only to these discrete time observations of $y_{t_k} = y_k$ and latent states $x_{t_k} = x_k$. Given the state of x_k, the distribution of observation y_k, also known as the emission probability, is known and the observations $y_k, k = 1, \ldots,$

are assumed independent conditionally on x_k. Furthermore, the latent process is assumed to be a Markov process, i.e., using the notation $x_{1:k} = (x_1, \ldots, x_k)$ then the *transition probability* of the latent process is $p(x_k|x_{1:k}) = p(x_k|x_{k-1})$. The dependence structure between x_k, y_k and x_{k-1} is visualized in Figure 3.3 where the arrows note the direction of the dependence, i.e., y_k depends on x_k, but not the other way around. The set of possible values of x_k is termed the state space, which we denote here as $\mathcal{X}$.

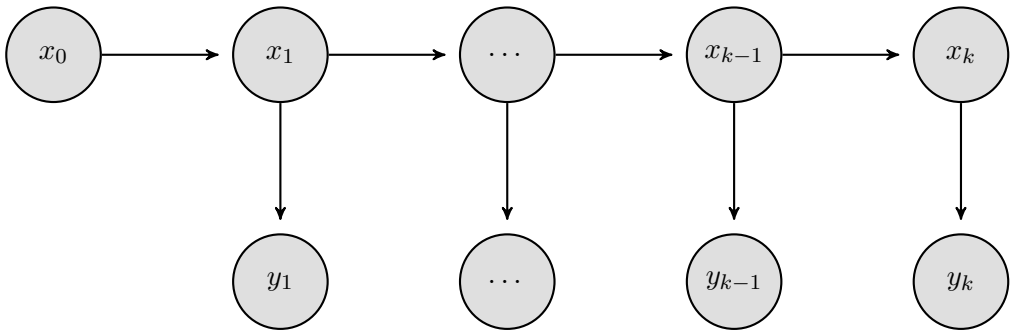

FIGURE 3.3: Graphical representation of a standard State Space Model with observation process y_k and latent process x_t.

Assume that the parameters θ is given, such that the model is fully specified. Also, given the transition probabilities for x_k along with either an initial distribution for x_0, $p(x_0)$, or x_0 itself, as well as the emission probabilities $p(y_k|x_k)$, then inference for a state space model amounts to answering the following question: *given observations $y_{1:s}$ what is the most probable state of x_k for $k = 0, \ldots, s$?* Depending on whether $s = k$ (filtering), $s > k$ (smoothing) or $s < k$ (forecasting), the problem is termed differently, but is solved similarly. The core of the issue is basically: we want to know something about x_k, which is unobserved and only known through the model design, by observing y_k. For a spike train model, we could let y_k denote the observed binary spike activity at time t_k and let x_k denote a process that influences the neurons behavior according to some dynamical evolution. Assuming for now that $k = s$, we can formulate the filtering problem using Bayes formula and exploiting the conditional independence of the observations y_k

$$\begin{aligned} p(x_k|y_{1:k}) &= \frac{p(x_k, y_{1:k})}{p(y_{1:k})} \\ &= \frac{p(y_{1:k}|x_k, y_{1:k-1})p(x_k|y_{1:k-1})}{p(y_k|y_{1:k-1})} \\ &= \frac{p(y_k|x_k)p(x_k|y_{1:k-1})}{p(y_k|y_{1:k-1})} \\ &\propto p(y_k|x_k)p(x_k|y_{1:k-1}). \end{aligned} \tag{3.15}$$

To see the proportionality in the last line, note that the denominator $p(y_k|y_{1:k-1})$ is redundant with regard to x_k, since it is simply a normalization constant and thus it does not include any information related to x_k. In the last line, the first probability is simply the observation likehood and the second is the one-step prediction of x_k given the past observations, also known as the (one-step) Chapman-Kolmogorov equation

$$p(x_k|y_{1:k-1}) = \int p(x_k|x_{k-1})p(x_{k-1}|y_{1:k-1})dx_{k-1}. \tag{3.16}$$

Starting with $p(x_0)$ and using (3.15) and (3.16) iteratively we end up with a recursive estimation scheme of the states $x_k, k = 0, 1, \ldots,$, alternating between predicting $p(x_k|y_{1:k-1})$ and updating $p(x_k|y_{1:k})$.

3.1 The Kalman Filter

While it is conceptually relatively straightforward to exploit the recursiveness of (3.15) and (3.16), evaluating and optimizing the prediction $p(x_k|y_{1:k-1})$ in practice is another thing entirely. Fortunately, for a particular subset of state space models, analytical solutions exist. If we assume that both x_k and y_k have linear Gaussian dynamics

$$\begin{aligned} x_k &= F_k x_{k-1} + w_k \\ y_k &= H_k x_k + v_k, \end{aligned} \tag{3.17}$$

where $w_k \sim \mathcal{N}(0, Q_k)$ and $v_k \sim \mathcal{N}(0, R_k)$ are both (multivariate) Gaussian with covariance Q_k and R_k respectively, then the Kalman filter provide an analytical solution to the filtering question, which is optimal in terms of the mean squared error. Since everything involves linear combinations of Gaussian densities, then both the prediction density for $p(x_k|y_{1:k-1})$ and the posterior density for $p(x_k|y_{1:k})$ will be Gaussian. Let $x_{k|k-1}$ and $W_{k|k-1}$ denote the mean and variance of the prediction density and $x_{k|k}$ and $W_{k|k}$ denote the mean and variance of the posterior density. The Kalman filter then consist of the following recursions

$$\text{predict:} \quad \begin{aligned} x_{k|k-1} &= F_k x_{k-1|k-1} \\ W_{k|k-1} &= F_k W_{k-1|k-1} F_k' + Q_k \end{aligned} \tag{3.18}$$

$$\text{update:} \quad \begin{aligned} x_{k|k} &= x_{k|k-1} + K_k\big(y_k - H_k x_{k|k-1}\big) \\ W_{k|k} &= W_{k|k-1} - K_k H_k W_{k|k-1}, \end{aligned} \tag{3.19}$$

where $K_k = W_{k|k-1} H_k' \big(H_k W_{k|k-1} H_k' + R_k\big)^{-1}$ is known as the *Kalman gain*. Equations (3.18) and (3.19) provide intuitive insight into the filtering mechanism. First a prediction is suggested by the model design (prediction from the prior) and secondly this prediction is corrected by the deviation from the actual observation (update of the posterior).

If the dynamics in (3.17) are nonlinear, but still Gaussian, i.e.

$$\begin{aligned} x_k &= f(x_{k-1}) + w_k \\ y_k &= g(x_k) + v_k, \end{aligned} \tag{3.20}$$

then a version of the Kalman filter known as the *extended Kalman filter* can be applied. This essentially works by linearizing the dynamics around the current estimates, by calculating the Jacobian matrices for f and g and evaluating at $x_{k|k}$ and y_k. These linear approximations are then used as substitutes for F_k and H_k in (3.17). However, due to these approximations, the extended Kalman filter does not ensure optimality of the solution, as was the case for the standard Kalman filter. Below we review how filtering can be solved for non-Gaussian processes and an analogue of the Kalman filter for point processes.

3.2 Stochastic State Point Process Filter

For non-Gaussian processes the Kalman filter cannot be directly applied, but in Eden et al. (2004) an analogue of the Kalman filter is derived for point processes observations by using a Gaussian approximation. The authors denote this the *stochastic state point process filter*. This can be applied to spike train observations y_k, with conditional intensity $\lambda(k|\ H_k, \theta_k)$, where $H_k = (y_{1:k-1}, \theta_{1:k-1}, x_{1:k-1})$ includes all relevant preceding history $y_{1:k-1}$ and a sequence of parameters θ_k as well as other auxiliary information (stimulus or other) in $x_{1:k-1}$. Here, the latent process is assumed to be the sequence of parameter estimates θ_k and as such the filter is used for parameter estimation for the point process model. The model dynamics for the latent state (parameter), analogue to (3.17), are assumed as

$$\theta_k = F_k\theta_k + w_k, \tag{3.21}$$

where $w_k \sim \mathcal{N}(0, Q_k)$ with Q_k as the covariance matrix. As in the case of the Kalman filter, the quest is to find the posterior density for the parameter θ_k

$$p(\theta_k|y_k, H_k) = \frac{p(y_k|\theta_k, H_k)p(\theta_k|H_k)}{p(y_k|\ H_k)} \tag{3.22}$$

Approximating (3.22) by a Gaussian density with mean $\theta_{k|k}$ and covariance $W_{k|k}$ implies that the one-step prediction

$$p(\theta_k|\ H_k) = \int p(\theta_k|\theta_{k-1}, H_k)p(\theta_{k-1}|y_{k-1}, H_{k-1})d\theta_{k-1}. \tag{3.23}$$

must be Gaussian, since the dynamics for θ in (3.21) were Gaussian, and thus (3.23) is a convolution of Gaussian densities and hence Gaussian itself. Denote by $\theta_{k|k-1}$ and $W_{k|k-1}$ the mean and covariance of the Gaussian density in (3.23) such that

$$p(\theta_k|\ H_k) \propto \exp\Big(-\frac{1}{2}(\theta_k - \theta_{k|k-1})'W_{k|k-1}^{-1}(\theta_k - \theta_{k|k-1})\Big).$$

Then from (3.22) we find that

$$\begin{aligned} p(\theta_k|y_k, H_k) &\propto p(y_k|\theta_k, H_k)\exp\Big(-\frac{1}{2}(\theta_k - \theta_{k|k-1})'W_{k|k-1}^{-1}(\theta_k - \theta_{k|k-1})\Big) \\ &\propto \exp\Big(-\frac{1}{2}(\theta_k - \theta_{k|k})'W_{k|k}^{-1}(\theta_k - \theta_{k|k})\Big), \end{aligned}$$

such that by applying the $\log$ at each side and using (3.7) to express the log-likelihood of the observation density $p(y_k|\theta_k, H_k)$, we end up with

$$\begin{aligned} &\log p(y_k|\theta_k, H_k) - \frac{1}{2}(\theta_k - \theta_{k|k-1})'W_{k|k-1}^{-1}(\theta_k - \theta_{k|k-1}) \\ &= y_k \log\lambda_k - \lambda_k\Delta t - \frac{1}{2}(\theta_k - \theta_{k|k-1})'W_{k|k-1}^{-1}(\theta_k - \theta_{k|k-1}) \\ &= -\frac{1}{2}(\theta_k - \theta_{k|k})'W_{k|k}^{-1}(\theta_k - \theta_{k|k}) + C, \end{aligned}$$

for some constant C and $\lambda_k = \lambda(k|\theta_k, H_k)$. Finally by differentiating with respect to θ_k, then

$$-\left(\frac{\partial \log \lambda_k}{\partial \theta_k}\right)'(y_k - \lambda_k \Delta t) + W_{k|k-1}^{-1}(\theta_k - \theta_{k|k-1}) = W_{k|k}^{-1}(\theta_k - \theta_{k|k}). \tag{3.24}$$

Now evaluating (3.24) at $\theta_k = \theta_{k|k-1}$ presents the equation

$$\left[\left(\frac{\partial \log \lambda_k}{\partial \theta_k}\right)'(y_k - \lambda_k \Delta t)\right]_{\theta_{k|k-1}} = -W_{k|k}^{-1}(\theta_{k|k-1} - \theta_{k|k}) \tag{3.25}$$

which in turn provide the update of the posterier mean

$$\theta_{k|k} = \theta_{k|k-1} + W_{k|k}\left[\left(\frac{\partial \log \lambda_k}{\partial \theta_k}\right)'(y_k - \lambda_k \Delta t)\right]_{\theta_{k|k-1}} \tag{3.26}$$

and by differentiating (3.24) and evaluating again at $\theta_k = \theta_{k|k-1}$ we find that

$$W_{k|k}^{-1} = W_{k|k-1}^{-1} + \left[\left(\frac{\partial \log \lambda_k}{\partial \theta_k}\right)'(\lambda_k \Delta t)\left(\frac{\partial \log \lambda_k}{\partial \theta_k'}\right) - (y_k - \lambda_k \Delta t)\left(\frac{\partial^2 \log \lambda_k}{\partial \theta_k \partial \theta_k'}\right)\right]_{\theta_{k|k-1}} \tag{3.27}$$

as the update equation for the posterior covariance. Putting all of it together, the algorithm is analogue to the Kalman filter where the prediction step is completely identical, due to (3.21), but the update is replaced by (3.26) and (3.27)

predict:
$$\begin{aligned} \theta_{k|k-1} &= F_k \theta_{k-1|k-1} \\ W_{k|k-1} &= F_k W_{k-1|k-1} F_k' + Q_k \end{aligned} \tag{3.28}$$

update:
$$\begin{aligned} W_{k|k}^{-1} &= W_{k|k-1}^{-1} + \left[\left(\frac{\partial \log \lambda_k}{\partial \theta_k}\right)'(\lambda_k \Delta t)\left(\frac{\partial \log \lambda_k}{\partial \theta_k'}\right) \right. \\ &\qquad \left. - (y_k - \lambda_k \Delta t)\left(\frac{\partial^2 \log \lambda_k}{\partial \theta_k \partial \theta_k'}\right)\right]_{\theta_{k|k-1}} \\ \theta_{k|k} &= \theta_{k|k-1} + W_{k|k}\left[\left(\frac{\partial \log \lambda_k}{\partial \theta_k}\right)'(y_k - \lambda_k \Delta t)\right]_{\theta_{k|k-1}}. \end{aligned} \tag{3.29}$$

3.3 Particle Filter for State Space Models

For general state space models with nonlinear dynamics

$$\begin{aligned} x_k &\sim p(x_k|x_{k-1}) \\ y_k &\sim p(y_k|x_k) \end{aligned} \tag{3.30}$$

and/or non-Gaussian distributions, the evaluation of (3.16) can be intractable and thus making it impossible or at least practically unfeasible to evaluate (3.15). In this case we need a different strategy to obtain an estimate of x_k.

For the general model (3.30) we can use a *particle filter* to approximate $p(x_k|y_{1:k})$ from a sample of M particles from $p(x_k|\ y_{1:k-1})$. Exploiting the technique of Monte Carlo integration then

$$p(x_k|y_{1:k}) \approx \frac{1}{M}\sum_{i=1}^{M}\delta_{(x_k)}(i), \tag{3.31}$$

where $\delta_{(x_t}(i)$ denote the Dirac measure and M is the number of particles. The intuition is, that if we sample enough particles from $p(x_k|y_{1:k})$ and use (3.31) for these, we effectively obtain a good estimate of the filtering density. However, sampling directly from $p(x_k|y_{1:k})$ is usually not feasible, or even possible. Hence, we need a way to obtain an approximate sample from $p(x_k|y_{1:k})$.

Consider (3.15) and (3.16) and note that in fact we need to approximate $p(x_k|y_{1:k-1})$ by the particles $\{x_k\}_{m=1}^{M}$. The strategy is then as follows: choose a tractable distribution to sample particles from, i.e. $x_k^m \sim q(x_k)$. The density $q(\cdot)$ is interchangeably termed the importance-, proposal- or instrumental density in the literature (Doucet and Johansen, 2009). The relation (3.31) then becomes

$$p(x_k|y_{1:k}) \approx \sum_{m=1}^{M} w_k^m x_k^m,$$

where

$$w_k^m \propto w_{k-1}^m p(y_k|x_k^m)\frac{p(x_k^m|y_{1:k-1})}{q(x_k^m)}. \tag{3.32}$$

However, choosing the transition density $q(x_k) = p(x_k|x_{k-1})$ as the importance density, then (3.32) simplifies to just evaluating the observation likelihood and normalizing the weights

$$\tilde{w}_k^m = w_{k-1}^m p(y_k|x_k^m)$$
$$w_k^m = w_{k-1}^m \frac{\tilde{w}_k^m}{\sum_{m=1}^{M}\tilde{w}_k^m}.$$

This version of a particle filter is termed the *bootstrap filter* and is straightforward to implement whenever it is feasible to sample directly from the transition density. However, since the most extreme particles, sampled in the tails of $p(x_k|x_{k-1})$ will have low weights, which are then carried over to the next time step by the scaling of w_{k-1}^m in (3.32), it is inevitable that these particles become degenerate. To overcome this problem a resampling step is often included in the algorithm (Cappé, Moulines, and Ryden, 2005). Basically this resets the weights of the particles to $1/M$, by resampling at time k using the weights $\{w_k^m\}_{m=1}^{M}$. This step removes most of the particles sampled in the tail, i.e., particles with low weights (and hence importance) in approximating the distribution, but duplicates particles that are in the regions of the density with more mass, and thus importance. Figure 3.4 shows a visual representation of the effects of resampling. There are various schemes for resampling, for instance multinomial sampling, but common to all is the computational cost of resampling. Therefore, if it is not necessary to resample at every step, a threshold can be used to determine when to resample. A simple choice is the *effective sample size*

(Cappé, Moulines, and Ryden, 2005)

$$N_{\text{eff}} = \left[\sum_{m=1}^{M} \left(\frac{w_k^m}{\sum_{n=1}^{M} w_k^n} \right)^2 \right]^{-1} \in [0, M],$$

which can be interpreted as the number of effective particles at time k. Thus, whenever $N_{\text{eff}} < cM$, for some scaling $c \in (0, 1)$, resampling is performed and the weights are reset such that $N_{\text{eff}} = M$.

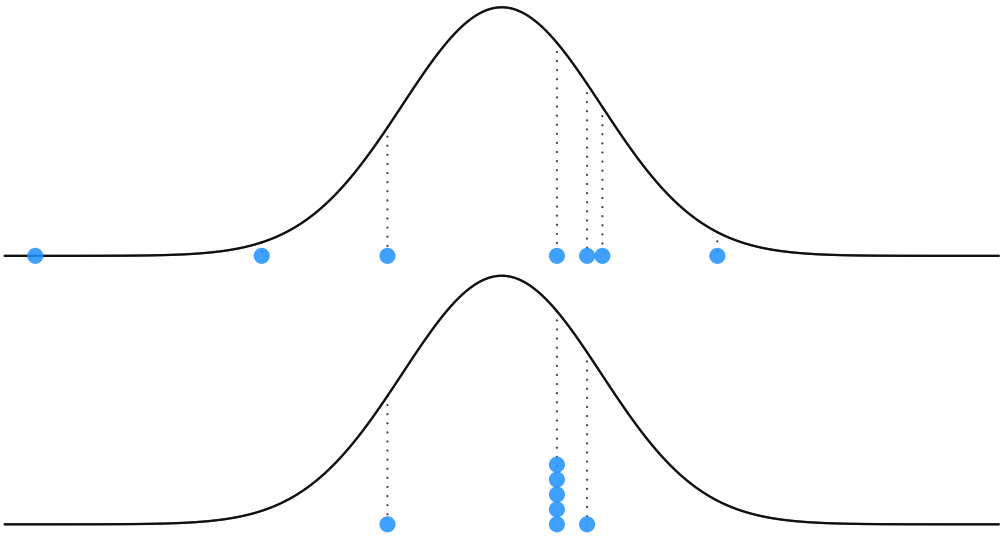

FIGURE 3.4: Resampling of particles. Top: Particles are sampled from a proposal distribution $q(x_k)$ and weighted according to the model density $p(x_k|x_{k-1})$. Bottom: After resampling, weights are uniform and particles with larger weights pre-resampling are duplicated.

The steps of the bootstrap filter algorithm, including resampling, is simply the following

1. Initialize particles at x_0 and sample $x_1^m \sim p(x_1|x_0)$ and evaluate $p(y_1|x_1^m)$.
2. For $k \geq 2$, sample $x_k \sim p(x_k|x_{k-1})$ and calculate weights (3.32).
3. If $N_{\text{eff}} < cM$, resample particles $x_{0:k}$ and reset weights $w_k^m = 1/M$ and continue from step 2.

3.4 Marginalized Particle Filter

When the state space model is nonlinear and at the same time high-dimensional, it becomes increasingly computationally expensive to obtain a good approximation of the posterior density from a particle filter, due to the increasing number of particles that are needed to keep up the quality of this approximation. However, if a subspace of the state space can be linearized conditionally on the remaining nonlinear subspace, one can exploit this linearization such that it is only necessary to use a particle filter on the remaining part of the state space. By conditioning on this particle approximation, the Kalman filter can then be applied for the linearized state space and thus simplifying calculations. This approach is termed the *marginalized particle filter* or *Rao-Blackwellized particle filter*.

In (Schon, Gustafsson, and Nordlund, 2005) the authors derive the marginalized filter by splitting the multi-dimensional state space x_k into the linear and non-linear parts

$$x_k = (x_k^{\text{lin}}, x_k^{non}).$$

The Kalman filter is applied to the linear state conditioned on the non-linear state, which is in turn handled by a particle filter, given by the relation

$$p(x_k|y_{1:k}) = p(x_k^{\text{lin}}, x_k^{\text{non}}|y_{1:k}) = \underbrace{p(x_k^{\text{lin}}|x_k^{\text{non}}, y_{1:k})}_{\text{Kalman Filter}} \underbrace{p(x_k^{\text{non}}|y_{1:k})}_{\text{Particle Filter}}. \tag{3.33}$$

By utilizing a combination of the Kalman filter, which is optimal for linear Gaussian state space models, and particle filter methods, the computational cost of a high-dimensional state space model is considerably reduced while the precision remain valid when linearization is a good approximation (see Schon, Gustafsson, and Nordlund, 2005). The marginalized particle filter algorithm is a revised version of the particle filter, where the Kalman optimization is included. Using the notation from Sections 3.1 and 3.3 it works as follows.

Initialize $x_{0|0}^{m,\text{lin}} = x_0$, $W_{0|0}^m = W_0$ for $m = 1, \ldots, M$ and sample particles $\{x_0^{m,\text{non}}\}_{m=1}^M$ from an initial distribution $x_0^{m,\text{non}} \sim p(x_0)$ and set all weights $w_0^m = M^{-1}$. Then for $k = 1, \ldots, K$ and each $m = 1, \ldots, M$ proceed with

1. Sample $x_k^{m,\text{non}} \sim p(x_k|x_{k-1}^{m,\text{non}}, y_{k-1})$.
2. Predict $x_{k|k-1}^{m,\text{lin}}$ and $W_{k|k-1}^m$ as in (3.18).
3. Update importance weights w_k^m according to (3.32) and normalize $\tilde{w}_k^m = \frac{w_k^m}{\sum_m w_k^m}$.
4. Resample the full particles $x_{1:k}^m = (x_{1:k|0:k-1}^{m,\text{lin}}, x_{1:k}^{m,\text{non}})$ according to the normalized weights $\tilde{w}_k^m$ and reset weights $w_k^m = M^{-1}$.
5. Update the linear posteriors $W_{k|k}^m$ and $x_{k|k}^{m,\text{lin}}$ using (3.19).
6. Set $k = k + 1$ and go to step 1.

MANUSCRIPT III

Capturing spike variability in noisy Izhikevich neurons using point process Generalized Linear Models

Jacob Østergaard, Mark Kramer and Uri Eden

Published in Neural Computation.
DOI: 10.1162/neco_a_01030

Abstract

To understand neural activity, two broad categories of models exist: statistical and dynamical. While statistical models possess rigorous methods for parameter estimation and goodness-of-fit assessment, dynamical models provide mechanistic insight. In general, these two categories of models are separately applied; understanding the relationships between these modeling approaches remains an area of active research. In this manuscript, we examine this relationship using simulation. To do so, we first generate spike train data from a well-known dynamical model - the Izhikevich neuron - with a noisy input current. We then fit these spike train data with a statistical model (a generalized linear model, or GLM with multiplicative influences of past spiking). For different levels of noise, we show how the GLM captures both the deterministic features of the Izhikevich neuron, and the variability driven by the noise. We conclude that the GLM captures essential features of the simulated spike trains, but for near deterministic spike trains, goodness-of-fit analyses reveal that the model does not fit very well in a statistical sense: the essential random part of the GLM is not captured.

1 Introduction

As recordings of neural activity become increasingly sophisticated, the resulting data become increasingly complex. Making sense of these data often requires more sophisticated approaches than visualization and simple summary statistics. One more advanced approach is the development and application of a model. Models serve both to characterize observed data and summarize the collective scientific

knowledge of the brain. In neuroscience - as in many other fields - these models are typically segregated into two categories: dynamical or mechanistic models, and statistical models. Dynamical models arise as an application of mathematical rules motivated by biophysical laws. These models tend to be deterministic - if not in practice, at least in spirit - and provide a mechanistic explanation for many dynamic brain activities. Statistical models are typically designed to capture data structure; these models often do not rely on neuronal biophysics.

The Generalized Linear Model (GLM) has been an essential part of modern statistics since its introduction by Nelder and Wedderburn (1972) and today these models are ubiquitous in statistical analyses in many differing fields. This includes biological processes, where in recent years GLMs have been used to describe coding properties and history dependence in neural spiking data (Kass and Ventura, 2001; Pillow et al., 2008; Sarma et al., 2012; Truccolo et al., 2005). The popularity of GLMs stems from the many advantageous features of this model class. The model defines a link function between the predictors and responses that ensures that the negative log likelihood is convex, allowing for fast optimal estimation of parameters, while still admitting flexible model capabilities, by utilizing general basis functions (McCullagh and Nelder, 1989). As with statistical models in general, the GLM quantifies both predictable and unpredictable structure in data and possesses efficient procedures for parameter estimation and model diagnostics and powerful tools to assess goodness-of-fit. An advantage of including uncertainty as an intrinsic model component in the design stage is that this random component can, to some extent, compensate for an incomplete specification of the factors influencing the observed data.

Dynamical models have been a pillar in data modeling for neuroscientists since Lapicque's 1907 integrate and fire model (Brunel and Rossum, 2007). Contrary to statistical models, these models typically focus on deterministic processes and are often built to explicitly represent mechanistic features of the data of interest. As such, dynamical models are often biophysically meaningful by construction, which may not be the case for a statistical model. Dynamical models are diverse, ranging from the biophysically realistic model of Hodgkin and Huxley (1952) to the more abstract model of Izhikevich (2003). The latter implements a simple mathematical design, while still producing realistic neural behaviors, such as bursting. Because of these properties, the Izhikevich model is a common choice, for example in large scale simulations of neurons with different behavior (Izhikevich and Edelman, 2008) and to simulate the impact of current input to large cortical networks (Ali, Sellers, and Fröhlich, 2013).

Historically, dynamical models have been used to examine deterministic or aggregate features of observed neuronal behavior. While dynamical models are often interpretable in terms of biophysical features, estimating parameters from individual spike trains is a complicated task. Most mechanistic models possess multiple parameters, which typically remain experimentally unconstrained. For example, the Hodgkin-Huxley model possesses at least seven parameters (e.g., capacitance, reversal potentials, and maximal conductances for the sodium, potassium, and leak currents) with wide ranges of possible values (Hodgkin and Huxley, 1952). In computational neuroscience, a common procedure for estimating these parameters is "hand-tuning" to produce simulated model dynamics that qualitatively match the desired neuronal activity (Prinz, Billimoria, and Marder, 2003). Hand-tuning approaches usually require a great deal of time and expertise (Traub et al., 1991; Nadim

et al., 1995; Vanier and Bower, 1999). Moreover, once a set of suitable parameters is found, it is often unclear whether the solution is unique or whether other model formulations exist that are compatible with the data (Prinz, Bucher, and Marder, 2004). On the other hand, statistical models are often used to estimate features and describe associations, but may not be directly physically interpretable.

Both modeling paradigms - dynamical and statistical - provide distinct advantages and disadvantages. Ideally, methods would exist to leverage the advantages of both appraoches, while mitigating their weaknesses. For example, a unified approach would allow researchers to better interpret statistical model results in terms of mechanistic features, while exploiting the mathematical theory of statistical models. This theory includes tools of parameter estimation, but also rigorous procedures to validate the model against data. The aim of this paper is to investigate how simple GLMs perform when used to describe the spiking patterns obtained from simulated Izhikevich neurons with noisy input. The noise is included to both resemble realistic physical observations, but more importantly to introduce and control variability of the data.

Previous work has analyzed the relationship between integrate-and-fire models and statistical techniques (Brunel and Latham, 2003; Paninski, 2006; Hertäg, Durstewitz, and Brunel, 2014). In these approaches, integrate-and-fire dynamical models were augmented by a stochastic component and a statistical approach was used to estimate the firing rate. Here we pursue a different strategy, starting with a well known, class of history dependent statistical models and analyzing the models' capabilities in capturing general features of various spike train patterns, generated from a dynamical model. The class of GLMs in this exposition is very flexible and, contrary to a dynamical model, there are no model assumptions regarding the specific mechanisms related to the physical processes that generate the spike train data, such as refractoriness. Recently, in Weber and Pillow (2017), the GLM framework has been shown to replicate multiple spiking patterns from the Izhikevich model, but to the authors' knowledge, no work has been presented that analyzes how well the GLM class captures features of different spike patterns under the varying influence of noise.

In this paper we show that the performance of the particular class of multiplicatively separable, history dependent GLMs varies with the amount of noise added to the input and with the large scale firing properties of the Izhikevich neuron. This class of GLMs is commonly used in spike train modeling (Pillow et al., 2008; Macke et al., 2011; Ahmadian, Pillow, and Paninski, 2011; Latimer et al., 2014). We examine how well the GLMs capture the intrinsic features of the individual neuron types, such as tonic spiking and bursting, as well as the variability in the individual spike trains originating from the noisy input. We first present the simulation model and define how variation is injected through both the input and the model parameters. Then the GLM design is outlined along with tools used for model assessment. In a series of numerical analyses, we demonstrate how GLMs behave for various levels of noise and different deterministic model features. We discuss in what ways the noise may affect interpretation of the GLM and the usage of this model class for spike train modeling by evaluating the goodness-of-fit of the models in terms of capturing both predictable and unpredictable structure. Finally, we highlight some of the shortcomings of simple multiplicatively separable GLMs and discuss extensions to the standard model that can mitigate some of the limitations identified in this paper.

2 Methods

2.1 Simulation of neuron activity

In this manuscript, we implement the Izhikevich neuron to simulate neural spiking activity (Izhikevich, 2003). The Izhikevich neuron is a relatively simple, dynamical model capable of reproducing many types of neural spiking behavior. We note that mechanistic models of neuron spiking can often be reduced to two-dimensional systems having a fast voltage variable and a slower recovery variable. For the neuron model considered here, we may interpret v as a voltage, and u as a recovery current, a phenomenological variable that represents the sum of all slow currents that modulate the spike generation mechanism (Izhikevich, 2010, chapter 8). The model includes 4 parameters a, b, c and d and a further input current I_t. Eq. (IV.20) shows the deterministic dynamical model for $t \geq 0$.

$$\begin{aligned} \frac{dv_t}{dt} &= 0.04v_t^2 + 5v_t + 140 - u_t + I_t \\ \frac{du_t}{dt} &= a(bv_t - u_t) \\ \text{if } v_t \geq 30 \text{ then } &\begin{cases} v_{t+} = c \\ u_{t+} = u_t + d. \end{cases} \end{aligned} \tag{III.1}$$

The dynamics in equation (IV.20) generates spikes whenever v_t passes a fixed threshold of 30. Immediately following a spike at time t (where $v_t \geq 30$), both variables are reset at time $t+$; v_t is reset to a fixed value given by the parameter c, and u_t is increased by an amount d. The parameters a, b, c, d determine the spike behavior; the parameters a and b act as a decay rate and a sensitivity parameter, respectively, and the parameters c and d determine how the variables are reset following a spike.

In order to introduce variation in simulations of (1), we let $I_t = I_0 + \sigma W_t$, where I_0, σ are constants and W_t denotes a standard white noise process (Øksendahl, 2007). The σ parameter controls the level of noise in the process, and in turn the variability of spiking behavior. We may interpret σ as the noise in the input driving the neuron. Another source of variability comes from the different parameter values used to simulate the spike trains. In this paper we model 6 types of neurons that produce distinct firing patterns (Izhikevich, 2003) when the input is constant. When $\sigma = 0$, the simulation model (IV.20) is deterministic and the neurons behave completely regularly, while for $\sigma > 0$ the model evolves stochastically. Assuming an Itô interpretation, model (IV.20) was simulated using a Euler-Maruyama scheme with timestep $\Delta t = 0.1$ ms, initial conditions $v_0 = -70, u_0 = bv_0$ and varying parameters a, b, c, d and I_0. We note that the 0.1 ms discretization is less than that used in Izhikevich (2003), and ensures that at most one spike occurs per discrete time bin. These simulations where then used to obtain spike trains. Although the Izhikevich neuron lacks the detailed biophysical mechanisms of more complex neuron models, this simple model serves as an excellent testbed for fitting the GLMs to diverse, realistic spike train behaviors, and exemplifying the issues in interpreting the results.

2.2 Model design

Let $N(t)$ denote the cumulative number of spikes observed up to time $t \in \mathbb{R}$. Define $\lambda(t|H_t, \theta)$ as the conditional intensity of $N(t)$, where H_t indicates the spiking history up to time t, and θ the model parameters. This conditional intensity defines the instantaneous probability of a spike given past spiking,

$$\lambda(t\big|H_t, \theta) = \lim_{\Delta t \to 0} \frac{P\big(\text{spike in } (t, t+\Delta t]\big|H_t, \theta\big)}{\Delta t}.$$

Given a set of observed spike times $\{s_j\}_{j=1}^n$ in an interval $[0, T]$, the log-likelihood

$$\log L(s_1, \ldots, s_n; \theta) = \int_0^T \log \lambda(t|H_t, \theta) dN(t) - \int_0^T \lambda(t|H_t, \theta) dt, \tag{III.2}$$

can be approximated as a discrete sum over individual time bins, by assuming that the number of spikes in each bin is Poisson distributed with a rate parameter that depends on past spiking for sufficiently small Δt, see Truccolo et al. (2005). The discrete likelihood approximation is

$$\log L(s_1, \ldots, s_n; \theta) \approx \sum_k \log \lambda(t_k|H_{t_k}, \theta) \Delta N(t_k) - \sum_k \lambda(t_k|H_{t_k}, \theta) \Delta t, \tag{III.3}$$

where $\Delta N(t_k)$ counts the spike activity in $(t_k, t_k + \Delta t]$. Since the Poisson and binomial distribution converge in the limit $\Delta t \to 0$, either could be assumed in the individual time bins. However, choosing the Poisson distribution leads to the canonical log-link function (McCullagh and Nelder, 1989).

The approximate log-likelihood (III.3) shows that choosing the form of the intensity $\lambda(t_k|H_{t_k}, \theta)$ is a crucial step of the model design. To include history dependence, we include a trailing history of the m past spike times, and in order to reduce the dimensionality of the model, we also introduce indicator basis functions that account for the spike windows in preceding time bins, wider than the observation bins of width Δt. In order to simplify the model formulation, we will assume that the influence of the previous spikes on the intensity is multiplicatively separable, which is a common model assumption when working with GLMs. Thus, when modelling the $\log$ intensity process we deal with a sum of components.

Defining the $(m \times p)$ matrix B with the (i, j)'th entry

$$b_{ij} = \mathbf{1}_{\big[(j-1)w+1, jw\big]}(i) = \begin{cases} 1 & \text{for } i \in \{(j-1)w+1, \ldots, jw\} \\ 0 & \text{else} \end{cases}, \tag{III.4}$$

where the parameter w denotes the number of Δt-wide time bins grouped, the model becomes

$$\log\Big(\lambda(t_k\big|H_{t_k})\Big) = \beta_0 + \sum_{j=1}^p \beta_j \sum_{i=1}^m y_{k-i} b_{ij}. \tag{III.5}$$

Here $y_k = \sum_{j=1}^n \mathbf{1}_{(t_k, t_k+\Delta t]}(s_j) = \Delta N(t_k)$ and the parameter vector is $\theta = (\beta_0, \beta_1, \ldots, \beta_p)$. With $\Delta t = 0.1$ ms and $w = 10$, each β_j in (III.5) relates to the total spiking activity in 1 ms time bins.

Defining the (exponential) filter $\mathcal{F}(\tau) = \exp\big(\sum_{j=1}^{p} \beta_j \mathbf{1}_{\left((j-1)w, jw\right]}(\tau)\big)$, this acts as the modulation of $\lambda_0 = \exp(\beta_0)$ at lagged time bins $t_{k-\tau}$.

As the randomness of the input vanishes, the y_k series becomes increasingly deterministic and tends to produce spiking at only a small set of interspike intervals. In this case only a few of the y_{k-i}'s in (5) are nonzero at regularly lagged time bins. Since $\lambda(t\big|H_t, \theta)\Delta t \approx P\big(\text{spike in } (t, t+\Delta t]\big|H_t, \theta\big)$, then $\lambda(t_k|H_{t_k}) = \exp\big(\beta_0 + \sum_{j=1}^{p} \beta_j \sum_{i=1}^{m} y_{k-i} b_{ij}\big) \to 0$ in intervals with no spike activity. Denoting the estimate of β_j by $\hat{\beta}_j$, this implies some estimates $\hat{\beta}_j$, must diverge to $-\infty$. On the other hand in bins of spike activity, $P\big(\text{spike in } (t, t+\Delta t]\big|H_t, \theta\big) \to 1$, when the spiking is deterministic. Hence, the estimated $\hat{\beta}_j$'s in these bins will amplify the model's probability of spiking in $(t, t+\Delta t]$ in order to match the observations and approximate the relation $\lambda(t|H_t) \to \frac{1}{\Delta t}$. To summarize, for a given interval $(t, t+\Delta t]$

$$\lim_{\Delta t \to 0} \lambda(t|H_t) = \begin{cases} \infty & \text{given spike activity} \\ 0 & \text{given no spike activity,} \end{cases}$$

implies that $\lim_{\Delta t \to 0} \lambda(t\big|H_t)$ describes a set of Dirac delta functions centered at the spike times. We note that an example of this finding will be illustrated in Figure III.4, which displays approximate delta functions at the periodic interval of spiking, when the data generating process is nearly deterministic.

2.3 Penalized GLM regression

Data with negligible variability can lead to convergence issues of parameter estimates in (III.5). If there is little or no variation in the data, the is a high chance of including perfect predictors in the model that will cause instability in the estimation of model parameters since the likelihood surface will be close to flat in certain directions, see Wedderburn (1976). Using a penalized regression is one way to handle these convergence issues and keep the parameter estimates finite.

Given a log-likelihood function $L(\theta)$ with parameters $\theta = (\beta_0, \beta_1, \ldots, \beta_p)$, a penalized regression will maximize $L(\theta)$ subject to a constraint

$$L_\kappa(\theta) = L(\theta) - \kappa \sum_{j=0}^{p} ||\beta_j||^q, \tag{III.6}$$

where $||\cdot||^q$ denotes the q-norm. For $q > 0$ the penalization in (6) will shrink the parameters towards 0. The choice of $q = 1$, corresponding to an L^1 penalization (LASSO regression) will, in addition to shrinkage, promote sparsity (Hastie, Tibshirani, and Friedman, 2001). The penalization parameter[1] κ determines how strong the shrinkage effect is, and $L_\kappa(\theta)$ implies that the optimization depends on some fixed κ. Choosing κ is a crucial part of using penalized regression, and therefore the choice should agree with the aim of the analysis. For the purposes presented in this exposition, we settled on two options that both depend on a goodness-of-fit test of an estimated model (see Section 2.5 below). The first choice was simply to chose κ such that the KS statistic was minimized; we will refer to this choice as *optimal KS*, where KS denotes the Kolmogorov-Smirnov test presented in Section 2.5 below. The

[1] We choose here to denote the penalization by κ, rather than the standard λ, in order not to confuse the penalization factor in the model fitting procedure with the intensity process, denoted $\lambda(\cdot)$.

other choice was to choose the maximum penalization, κ, such that the goodness-of-fit p-value was insignificant with respect to a predetermined threshold denoting the significance level. In what follows, we chose a threshold significance level of 0.05. We will refer to the second choice for κ as *maximum* κ. Obviously the two differ in behavior. The first choice of κ selects models with a better goodness-of-fit, whereas the second selects more sparse models due to the choice of $L1$ penalization in (III.6). We use the second choice (maximum κ) as the default to impose sparsity. However, to illustrate certain points, we also refer to models chosen by optimal KS in some parts. In the case that no p-values were above the threshold, or if an optimal value could not be determined, the value of κ was set as the smallest values tried in the model estimation procedure, i.e. $\kappa \approx 0$. This occurred for 66% of the simulations for non-bursting neurons, and 12% of bursting neurons. We note that $\kappa \approx 0$ still includes a small penalty so that the divergence of β_j estimates is avoided, but less parameters are set to zero with the LASSO regression.

2.4 Approximate covariance matrix

The observed Fisher information matrix $\widehat{\mathbf{F}}$ was used to estimate the covariance matrix for the penalized GLM. The observed Fisher information for an unrestricted GLM model is given as

$$\widehat{\mathbf{F}} = X'WX,$$

where X is the so called "design matrix" of the GLM that contain the predictors of the model and $\hat{W}$ is a model specific diagonal matrix that is computed as part of the estimation procedure (McCullagh and Nelder, 1989). For the models considered here, $X = HB$, where B is the matrix of indicator basis vectors described in Section 2.2, and H is the matrix of m lagged observations, such that the k^{th} row of H is $H_{(k)} = (y_{k-1}, \ldots, y_{k-m})$. For a Poisson GLM, the diagonal matrix $\hat{W}$ consists of the predicted intensities $\hat{\lambda}(t_k \big| H_{t_k})$. The inverse of the observed Fisher information, $\widehat{\mathbf{F}}^{-1}$, was then used as an estimate of the covariance matrix of $\hat{\beta}$. Due to the penalization in (III.6), parameters were restricted from converging toward $\pm\infty$, but because of this convergence issue, the unrestricted model would not admit a sensible covariance estimates. The estimate $\widehat{\mathbf{F}}^{-1}$ does not account for a penalization of the likelihood, and therefore the matrix $\widehat{\mathrm{Cov}}(\hat{\beta}) = \widehat{\mathbf{F}}^{-1}$ should only be interpreted as a rough approximation for the penalized model. Recent work addresses the bias in the covariance estimator due to inclusion of a penalization term in the model formulation (Geer et al., 2014; Taylor and Tibshirani, 2015); however, standard software packages for LASSO estimation (e.g., `glmnet`, used here (Friedman, Hastie, and Tibshirani, 2010)) have not yet implemented a correction for this bias. In this manuscript, we focus on the sign of the covariance estimate and not the specific value. Thus, although biased, $\widehat{\mathbf{F}}^{-1}$ can still reveal trends in the parameters. Defining $\Lambda = \sqrt{\mathrm{diag}(\widehat{\mathbf{F}}^{-1})}$, then

$$\widehat{\mathrm{Cor}}(\hat{\beta}) = \Lambda^{-1}\widehat{\mathbf{F}}^{-1}\Lambda^{-1} \tag{III.7}$$

is an estimate of the correlation matrix for $\hat{\beta}$. In the results section we present correlation matrix estimates based on (III.7).

2.5 Goodness-of-fit

In order to evaluate a statistical model, one must take into account both the structural and the random components of the model. Intuitively one can think of this as measuring the model's ability to capture both the structure of the observed data (structural component) as well as generalizability, by compensating for features that are not accounted for in the structural component through the statistical distribution (random component). Thus, with deterministic input, it is possible that a statistical model can describe the observed features to near perfection. However, such a model will rarely predict the features of some other data with only slight variations to the first. This lack of generalizability, caused by a inadequate fit of the random component in deterministic settings, will lead to poor statistical model diagnostics.

To assess the models' goodness-of-fit, the time rescaling theorem (Brown et al., 2002), was applied to the observed spike times $\{s_j\}_{j=1}^n$, using the estimated intensity process $\hat{\lambda}$, in order to obtain rescaled spike times $\{z_j\}_{j=1}^n$. The empirical distribution of the z_j's was then compared to the theoretical $\mathrm{Exp}(1)$ distribution using the Kolmogorov-Smirnov (KS) test statistic (Kass, Eden, and Brown, 2014). This statistic is given by

$$D_n = \sup |\hat{F}_n(x) - F_0(x)|, \tag{III.8}$$

where $\hat{F}_n(x)$ and $F_0(x)$ denote the empirical and theoretical cumulative distribution functions (CDFs) respectively. Plotting $\hat{F}_n(x)$ against $F_0(x)$ with approximate 95% confidence bounds $\pm\frac{1.36}{\sqrt{n}}$, as suggested by Kass, Eden, and Brown (2014), provides a visual assessment of the KS test. For a detailed discussion of these goodness-of-fit procedures, we refer the reader to Brown et al. (2002) and Kass, Eden, and Brown (2014).

In addition to the KS statistic (III.8), the relative deviance was used to evaluate the model fit. While the statistic D_n measures how the empirical model deviates from a theoretical model, the relative deviance can reveal the amount of structure that is present in the data. The relative deviance measures how an estimated model captures data features. The "smallest" model deviance for a given data set occurs with the saturated model, where the number of parameters equals the number of data points. Due to the equal number of parameters and data points, there is no data left to estimate variability of the data. Therefore, the saturated model will completely describe the observed data set, but will not generalize well to another dataset. As such, we could say that the saturated model is the "most structural" we can define for a specific data set, as it is purely descriptive. The other extreme is the null model, which is the 1 parameter model including an intercept, β_0, only. For the null model, only a single parameter is estimated, in this case defining a homogeneous Poisson process with no influence of past spiking. It is thus an example of the "most stochastic" model we can define, in the sense that it maximizes the entropy of the spike counting process. The relative deviance measures where an estimated model lies in the spectrum between these two extremes, with a relative deviance of 0 corresponding to a fit equal to the saturated model and a relative deviance of 1 corresponding to a fit equal to the null model.

It is important to note that for a traditional deviance analysis, the goodness-of-fit of multiple models is compared by assessing their deviance on the same dataset.

Here, we examine for one particular class of GLMs the relative deviance across different datasets with different levels of input current noise. As such, a lower relative deviance of a different noise level should not be interpreted as an improved fit, but as one whose description of the data structure is closer to that of a saturated model.

3 Results

Spike trains for six types of Izhikevich neurons were simulated with parameters displayed in Table III.1. We note that the six types of neurons considered here mimic spiking behavior observed in *in vivo* and *in vitro* neural recordings (Izhikevich, 2004).

Neuron Type	a	b	c	d	I_0
Tonic Spiking	0.02	0.20	-65	6	14
Phasic Spiking	0.02	0.25	-65	6	1
Tonic Bursting	0.02	0.20	-50	2	10
Phasic Bursting	0.02	0.25	-55	0.05	1
Mixed Mode	0.02	0.20	-55	4	10
Spike Frequency Adaptation	0.01	0.20	-65	8	20

TABLE III.1: Model parameters a, b, c, d used to simulate the 6 types of neurons and the mean input I_0. Parameter values and neuron type from Izhikevich (2004).

For each type of neuron, 10 spike trains were simulated for 29 values of σ evenly distributed in the interval $(0, 20]$. Hence, the simulations ranged from almost deterministic $(\sigma = 0.1)$ to almost completely random spiking ($\sigma = 20$). As σ is increased, the intrinsic features of the individual neuron types become progressively noise driven and indistinguishable. When $\sigma >> 0$ the spiking activity is based more on the randomness of the input than the model parameters. Each simulation consisted of $2 \cdot 10^5$ observations at times $t_k, k = 1, \ldots, 2 \cdot 10^5$, with timestep $\Delta t = 0.1$ ms, corresponding to 20 seconds of observations for each simulation. The spike trains were derived from the simulated voltage trajectories by determining when $v_{t_k} \geq 30$. Figure III.1 presents the simulated spike trains for each type of neuron for varying σ values. An L^1 penalized regression for a Poisson GLM of the form (III.5) was fitted for each simulated spike train, where the value of κ in (III.6) was based on either optimal KS or maximum κ (see Section 2.3), depending on the analysis.

The history dependence of the GLM was set to 100 ms, corresponding to $m = 1000$ timesteps for a discretization step of $\Delta t = 0.1$ ms. With the indicator width set to 1 ms ($w = 10$), the model (III.5) had a maximum of 100+1 parameters. Some of the parameters are shrunk to zero by the penalized regression procedure, and as such the resulting filters consisted of less than 101 nonzero parameters.

The numerical analysis was carried out with the statistical programming environment *R*, R Core Team (2017), using the package `glmnet` to perform penalized regression for GLMs.

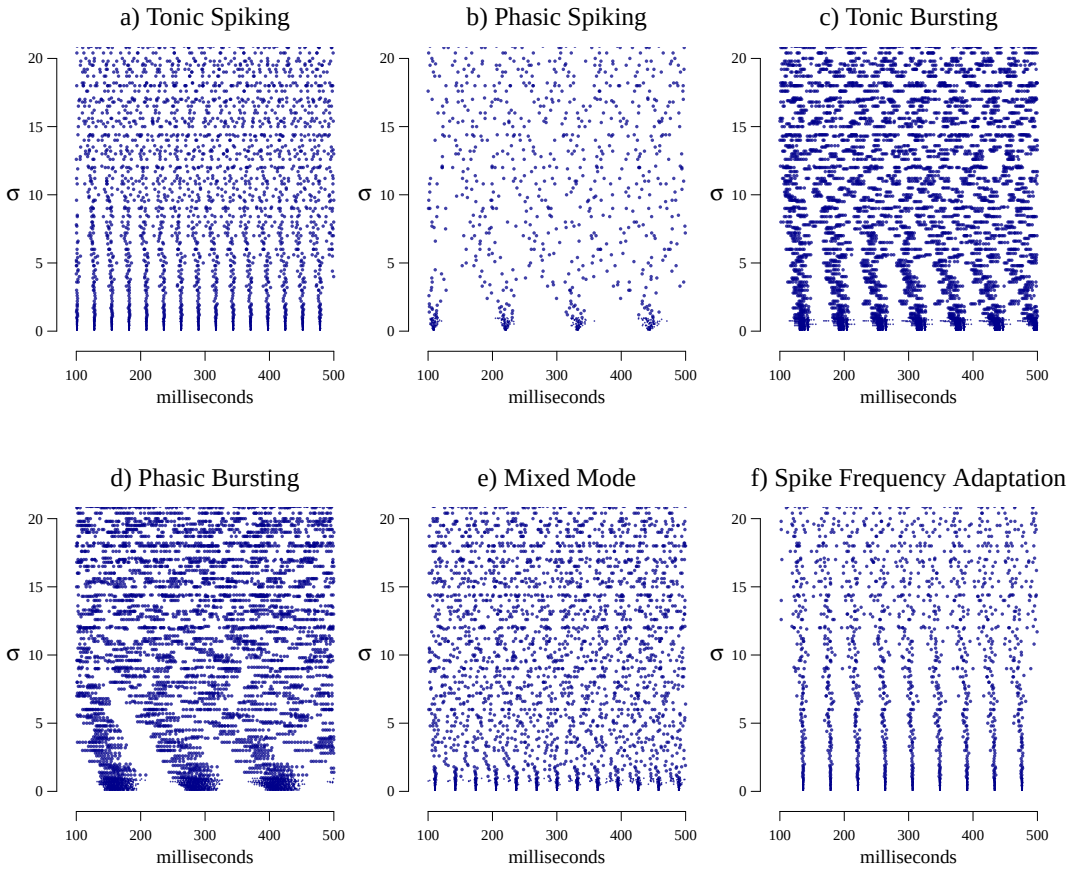

FIGURE III.1: **Simulated spike trains for various neuron types and values of σ.** Each type displays nearly regular behavior as σ approaches 0 and almost completely random patterns as σ approaches 20. Distortion of the individual neuron type behavior varies from rapid (Phasic Spiking and Mixed Mode) to gradual (Tonic Spiking and Spike Frequency Adaption). Burst periods remain present at high levels of noise, and the Spike Frequency Adaption neuron retains features of its spike pattern for σ as high as 15.

3.1 Tonic spiking and bursting with intermediate noise

We first present a more detailed analysis of the Tonic Spiking and Tonic Bursting neurons for a single value of σ to illustrate the model fitting and the goodness-of-fit analyses. These neuron types were chosen specifically to investigate how the GLM handles the regular behavior of a tonic spiking neuron, which possess a uni-modal interspike-interval (ISI) distribution, versus the switching behavior of a bursting neuron, which produces a bi-modal ISI distribution. The value of $\sigma = 5$ was chosen such that the neurons display both predictable structure of interest as well as variability, as evident from Figures III.1a and III.1c. The analyses for the two types of neurons are presented in Figures III.2 and III.3 for the Tonic Spiking and Tonic Bursting neurons, respectively. These figures display an interval of simulated spiking activity for the neuron, with the estimated intensity function $\hat{\lambda}(t|H_t)$ below, the corresponding histogram of ISIs, the histogram of rescaled spike times, the estimated filter, a KS plot of $\hat{F}_n(x)$ vs. $F_0(x)$ with approximate 95% confidence bounds, and a plot of the residual process $N(t) - \int_0^t \hat{\lambda}(t|H_t)dt$. Penalization was set by maximum κ with KS statistic significance threshold above 0.05 for both analyses.

The estimated intensity for the Tonic Spiking neuron (Figure III.2b) displays peaks that coincide with the spikes of the potential (Figure III.2a). There is a rise in the intensity prior to spiking and a instantaneous drop immediately following an observed spike, which implies that the model captures the predictable structure well. The ISI histogram for the Tonic Spiking neuron (Figure III.2d) is unimodal with a mean of 26.6 ms, and the width of the histogram implies some variability between spike times, which is expected for the choice of $\sigma = 5$. The estimated filter (Figure III.2e) displays a refractory period up to 20 ms and shows peaks around multiples of 26.6 ms, indicating that the model captures the regularity of spiking; this result agrees well with the type of neuron considered. The multiple peaks in the filter suggest that at this level of σ, the current probability of spiking is well predicted by the most recent spike as well as the previous two spikes. The rescaled spike time histogram (Figure III.2f), overlaid with an $\mathrm{Exp}(1)$ distribution (red line), shows that the model agrees well with the theoretical distribution, which is further supported by the KS plot (Figure III.2g). Furthermore, the KS statistic, $D_n^{\text{spiking}} = 0.046$, yields a p-value of 0.088, in accordance with the choice of κ such that the p-value is > 0.05. For reference, using optimal KS instead to set κ, the KS statistic was 0.022 with a p-value of 0.88. Inspection of the residuals (first quarter of the total residual process is shown in Figure III.2c) does not give rise to concerns regarding time-dependent trends.

For the Tonic Bursting model in Figure III.3, the estimated intensity (Figure III.3b) shows that bursting is well captured; we note that increases in the estimated intensity occur just before the times of action potential bursts (Figure III.3a). The rise of the intensity before a burst varies slightly between individual bursts and the drop after a period of bursting is not as sharp as for the case for Tonic Spiking. These observations suggest that it is more complicated to capture the beginnings and ends of bursts than single spikes. We note that the filter (Figure III.3e) does not possess a second peak between 40-50 ms, as seen in the ISI histogram (Figure III.3d). Instead, there is an inflection in the inhibitory trough of the filter starting around 40 ms. When convolved with the past bursting activity this inflection leads to a rapid rise in intensity approximately 40 ms after the final spike in the previous burst. The KS

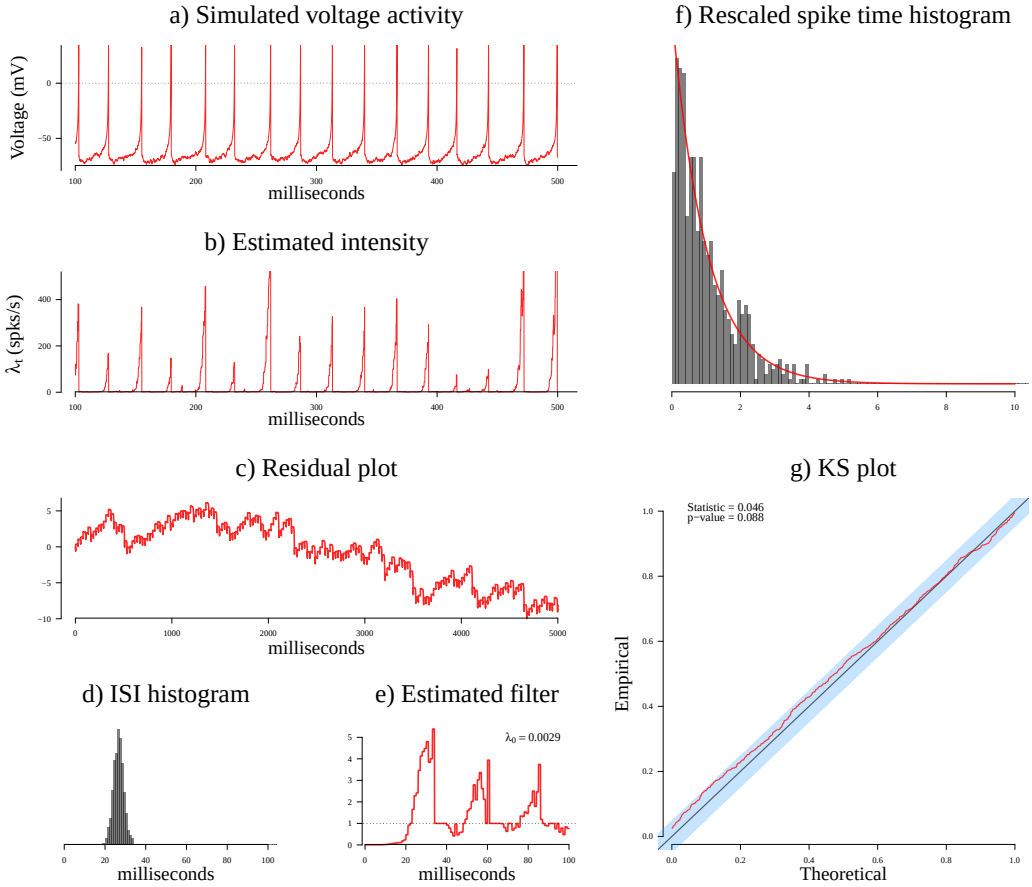

FIGURE III.2: **At a moderate level of noise, the GLM captures features of the Tonic Spiking neuron.** Example of the simulated voltage activity (a) and the corresponding estimated intensity process (b). The residual plot (c) does not show any trending behavior. The ISI histogram (d) possesses a broad peak and the peaks in the estimated filter (e) are consistent with the approximate interval between spikes. The rescaled spike time histogram (f) is well approximated by an Exp(1) probability density function (red curve) and the KS plot (g) indicates a decent fit.

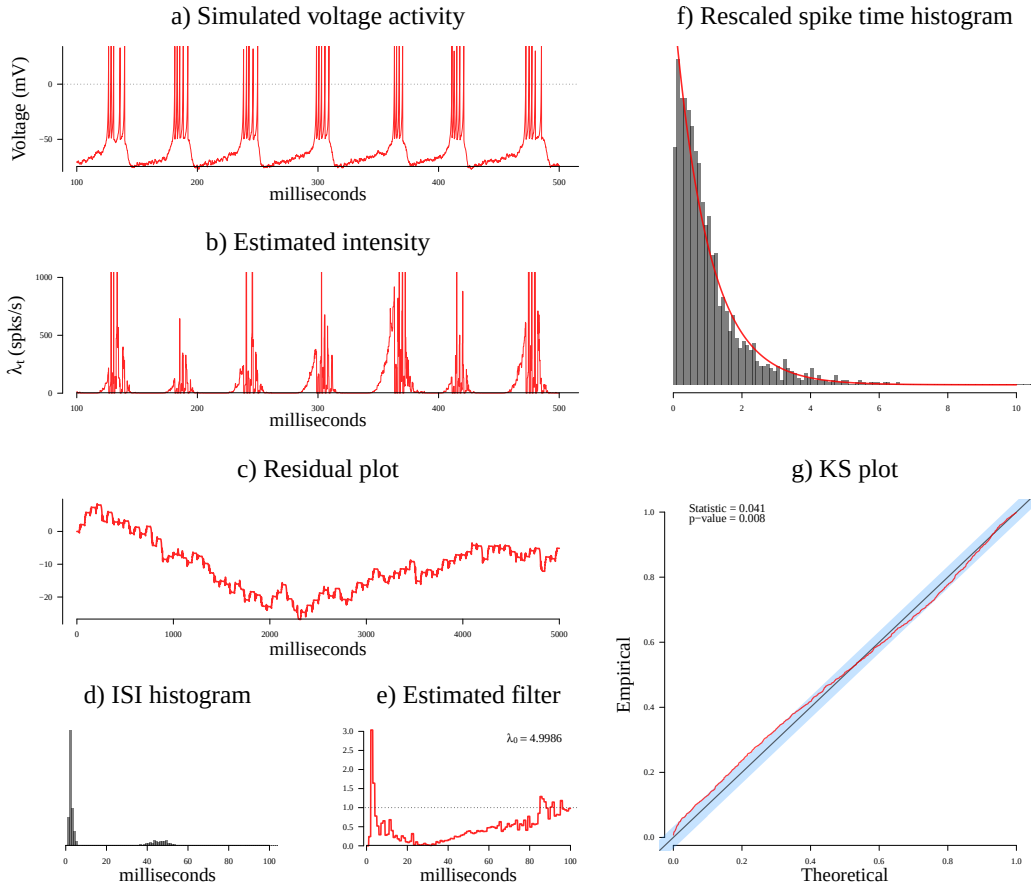

FIGURE III.3: **At a moderate level of noise, the GLM captures features of the Tonic Bursting neuron.** Example of the simulated voltage activity (a) and the corresponding estimated intensity process (b). The residual process (c) does not display trends, and the bi-modal ISI histogram (d) is approximated by the filter (e). The rescaled spike time histogram (f) is well approximated by an Exp(1) probability density function (red curve), while the KS plot (g) shows slight overestimation of short ISIs and underestimation of longer ISIs.

plot in Figure III.3g and the residual process in Figure III.3c provide further evidence that the GLM captures the structure in the bursting data well.

The rescaled ISI histogram in Figure III.3d agrees with the $\mathrm{Exp}(1)$ distribution, and the KS statistic is $D_n^{\text{burst}} = 0.041$, which is less than the value for the Tonic Spiking neuron: $D_n^{\text{spiking}} = 0.046$. However, the corresponding p-value is significantly lower at 0.008, because for the Tonic Bursting neuron, the number of observed spikes is $N^{\text{bursting}}(T) = 1668$ compared to $N^{\text{spiking}}(T) = 757$ for the Tonic Spiking neuron. The larger number of spikes results in a much lower p-value for Tonic Bursting, since the KS test is sensitive to the number of observed spikes. Using optimal KS to set κ for the Tonic Bursting did not improve the model fit; the KS statistic was 0.040 with a p-value of 0.009.

3.2 Spike history filters

We now turn to a more general analysis of the models, across a wider range of firing properties and noise levels, σ. Figure III.4 presents the estimated GLM filters for combinations of simulated neuron types and σ values. The penalization parameter κ was set by maximum κ with a threshold of 0.05; see Section 2.3. Filters for regularly spiking neurons, such as the Tonic Spiking neuron, Figure III.4a, and the Spike Frequency Adaption neuron, Figure III.4f, display peaks in the filters for low to intermediate values of σ. For $\sigma \to 0$ the filters approach collections of delta functions, with mass concentrated around multiples of the deterministic (ISI) period. For example, for the deterministic Tonic Spiking neuron, the ISI is ≈ 27 ms, hence the peaks in the filter $\mathcal{F}$ appear near 27, 54 and 81 ms in agreement with the regular ISIs of the Tonic Spiking neuron. For the bursting neurons, Figures III.4c and III.4d, there is a narrow peak approximately at 2-5 ms, except for very small σ. Decreases in the filter follow the peaks at short times, indicating that refractory periods are captured by the model, for all combinations of type and σ. However, the refractory period varies as σ increases: either increasing (for Tonic Spiking) or decreasing (Phasic Spiking), or remaining approximately unchanged (for Tonic Bursting). This indicates a minimal absolute refractory period for bursting neurons, that is independent of the input noise. Common to all filters, is a flattening of $\mathcal{F}$ as σ increases, most visibly for regularly spiking neurons (Tonic Spiking and Spike Frequency Adaption).

This implies that, by increasing the injected noise σ, the spiking approaches a homogeneous Poisson process and it becomes progressively difficult to extract mechanistic structure, such as the interval of regular spiking (Figure III.4a) or inter-bursting activity.

3.3 Goodness-of-Fit

KS tests were performed for combinations of neuron type and σ to assess each model's goodness-of-fit. Figure III.5 presents an overview of these results. The plot displays the $-\log(D_n)$, where D_n is the KS statistic (III.8), and the relative deviance respectively. The statistics, D_n, were log-transformed to emphasize trends as σ varies. Furthermore, in order to present implications of changing σ, the penalization parameter κ was set by optimal KS, such that different σ's could lead to different penalizations. Had we instead chosen penalization by maximum κ, the KS statistics would not reveal an optimal range of σ.

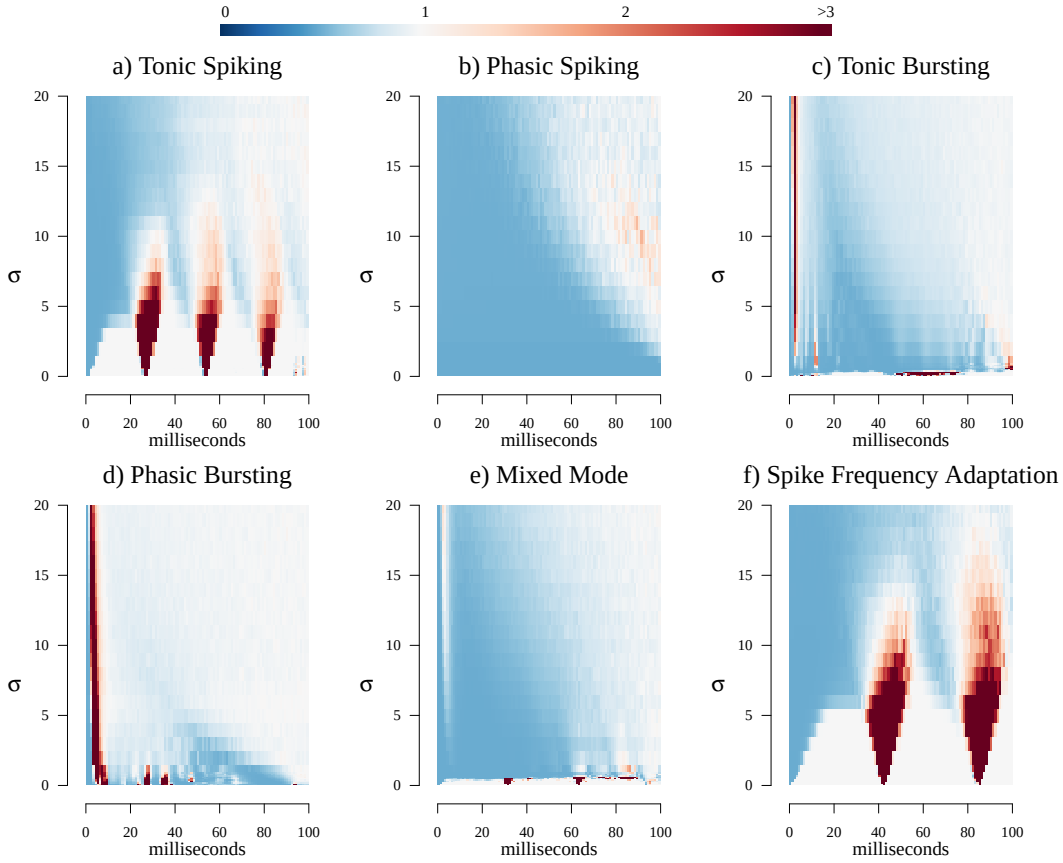

FIGURE III.4: **Estimated filters $\mathcal{F}$ using maximum κ penalization with threshold 0.05, as functions of lagged time (in ms) and σ.** Blue colors indicate refractory periods ($\mathcal{F} < 1$) and red colors indicate excitatory effects ($\mathcal{F} > 1$). White indicates $\mathcal{F} \approx 1$. Peaks are clear for tonic spiking and spike frequency adaption neurons in (a) and (f), as well as for bursting neurons in (c) and (d). Refractory periods are present for all types and vary with σ. Note that the $\mathcal{F}$'s converge towards collections of delta functions as σ approaches 0.

The $-\log(D_n)$ for all simulated neuron types in Figure III.5 increases quickly as σ increases from 0 to 2, suggesting that for low values of σ, the GLM has difficulty in capturing the near deterministic spiking distribution, but does well at capturing this distribution once there is some variability present. The corresponding p-values are above 0.05 for $\sigma > 1$ for all neuron types, except Phasic Bursting and Tonic Bursting. For Phasic Bursting, the p-values exceed 0.05 for $\sigma \in [5, 6]$, where as for Tonic Bursting the p-values never exceed 0.05. Except for Mixed Mode and Spike Frequency Adaption, all types display optimal values in a range of σ's, where as the mentioned types seem to settle at a plateau for $\sigma \geq 4$. The optimal range, both the width and location, of σ depends on the type of neuron.

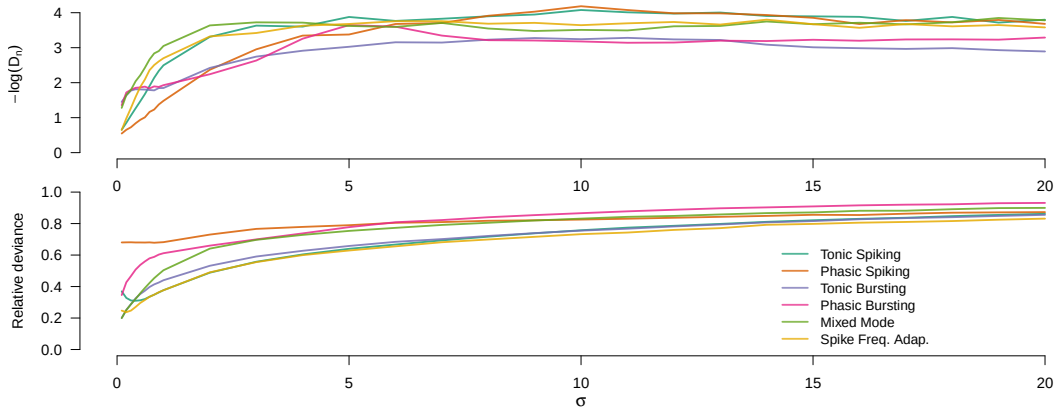

FIGURE III.5: **Two methods to assess goodness-of-fit for varying σ.** Top: For the log-transformed KS statistics D_n, larger values correspond to smaller KS statistics. Bottom: Relative deviance of the estimated model. A value of 0 indicates a fit equal to the saturated model, and a value of 1 indicates a fit equal to the null model.

For the relative deviances, we chose to use maximum κ. The increasing trend in the relative deviance suggests that, for low values of σ, there is more predictable structure captured by the model, whereas for higher noise levels, the estimated model approaches a homogeneous Poisson model, which includes only a baseline firing parameter. This general trend is consistent with our intuition for the model behavior; as the randomness of the input overshadows features of the Izhikevich model dynamics, the spike trains become better described by a homogeneous Poisson process.

We conclude from these two assessments that an optimal range for σ exists in which the GLM captures both the variability and predictable structure of the spike trains simulated using the noisy Izhikevich neurons. For very low values of σ, there is little variability to be captured by the GLM, while for very high values of σ, there is little predictable structure beyond the baseline firing rate. However, a general optimal range of σ cannot be explicitly defined as it depends on the type of neuron. Thus, when analyzing data, multiple goodness-of-fit measures should be used to assess the degree to which the model fits the variability (eg. KS test) and the predictable structure in the model (eg. relative deviance).

3.4 Model structure for extreme σ values

The results in Figure III.4 display the estimated filter parameters as $\sigma \to 0$. We note that the maximum estimated filter values are capped at a value of 3 in Figure III.4, so that the full extent of the peak values is not directly visualized. The peaks, increasing in height and decreasing in width, resemble delta functions as the noise level approach 0. In the other extreme of large σ, most of the visible structure in the filter vanishes, except for a refractory period. Except in the case of bursting neurons, the filters flatten (i.e., approach 1) for large σ, implying that less neuron specific structure is captured by the models. This finding is consistent with the results in Figure III.5, where the relative deviances approach 1 (the null model) with low values of the corresponding KS statistics, indicating that the simulated spike trains are well approximated by a GLM with an intercept (baseline) term only, which corresponds to a homogeneous Poisson process.

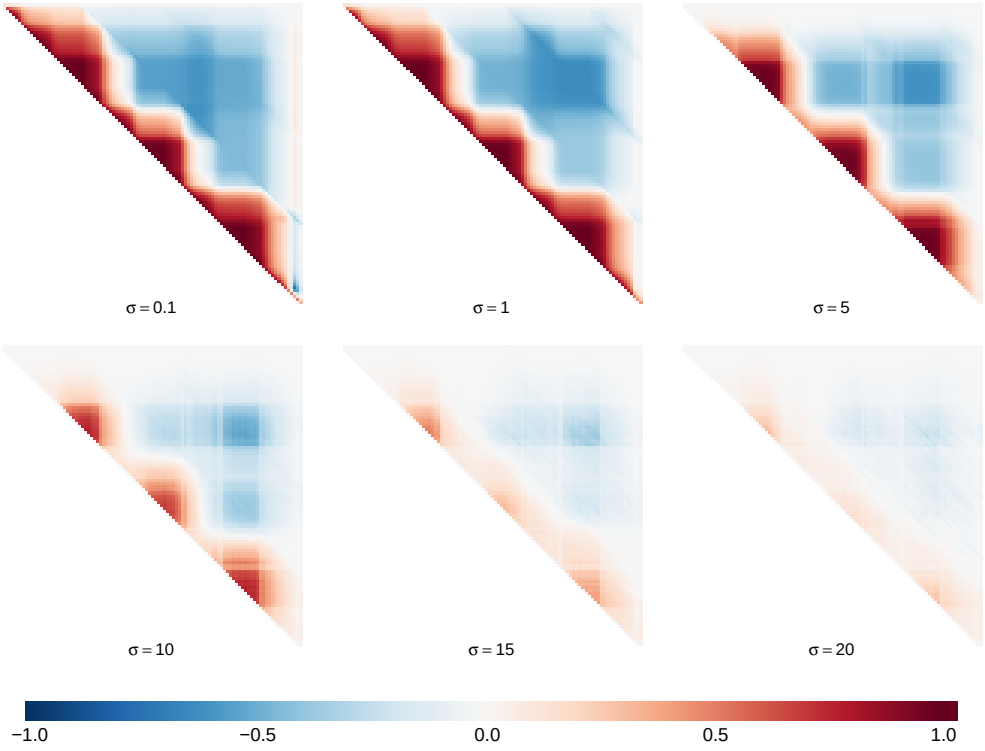

FIGURE III.6: **Estimated correlation matrices of the Tonic Spiking neuron with $\sigma = 0.1, 1, 5, 10, 15, 20$.** Notice how the correlation structure vanishes as σ increases. The case $\sigma = 5$ (top, right) displays three noticeable bumps, corresponding to the three peaks of the filter in Figure III.2. There are clear positive correlation among lags close to each other, and negative correlation among lags further apart.

The estimated correlation matrices (see Section 2.4) for the Tonic Spiking neuron in Figure III.6, reveal that for low σ values there is positive correlation (red) between parameters that are close to each other in the temporal dimension along the diagonal, but at the three peaks the positive correlation extends further in time. There is negative correlation (blue) among variables that are far apart in the temporal dimension, but most interestingly the clusters of variables associated with each peak exhibit negative correlation.

The negative correlation between peak variables suggests that past spiking at these lags is highly correlated and that, as predictors they are largely redundant. We note that the neuron spike probability can be impacted by the first, second or third last recorded spike, but it is the total accumulated effect of the past 100 ms that determines the current spike probability. As such, the filter should be evaluated as a whole, rather than at individual points in time.

As $\sigma > 5$, the correlation structure starts to disappear and for higher values of σ the correlation structure in the model is greatly reduced. Once more, this is in line with the findings in Figure III.5 as models for very noisy data capture mostly baseline activity.

For the Tonic Bursting neuron, the correlation matrices (not shown) showed little or no structure beyond a few milliseconds off-diagonal. Based on Figure III.4, this lack of correlation structure is not surprising. For higher values of σ the filter only captured the increased probability of a spike with short delay, but as the remaining part of the filter approaches 1, the (irregular) inter-burst periods were not captured. Thus, any relatively strong dependence beyond the intra-burst interval is not expected as σ increases. As in the case of the Tonic Spiking filters, the entire filter should be interpreted for Tonic Bursting. Thus, it is not only the time of the last spike that determines the current probability, but rather the 100 ms history of spikes, which in the case of bursting can include multiple spikes in a burst or multiple burst periods, that impacts the probability of spiking.

4 Discussion

In this paper we have examined how well a commonly used class of GLMs performs, when used to capture both predictable structure and variability of spike trains derived from simulated, noisy Izhikevich neurons. A useful model is one that allows for clear, simple interpretations, which depends critically on the form of the model. When referring to GLMs in this paper, we implicitly mean the specific class of multiplicatively separable history dependent GLMs. These GLMs were designed to capture the influence of past spiking on the current firing intensity using a simple set of indicator basis functions, and assuming that the influences of previous spikes are multiplicatively separable. This indicator basis is well suited to display the approximation of delta functions, when the noise level converges to 0 and the spike trains become increasingly deterministic. However, using indicator basis functions increases the number of parameters compared to other choices such as splines bases, and can lead to the problem of perfect separation. Therefore, it was necessary to use a penalized regression to fit the models with indicator basis functions for low noise levels where only a few parameters suffice to describe the structure. For this reason, we opted for an L^1 penalization, corresponding to LASSO regression, which shrinks redundant parameters to 0 by promoting sparsity. The implementation of the LASSO regression we used here utilized cyclical coordinate descent (Friedman, Hastie, and Tibshirani, 2010), when optimizing the likelihood, where as the standard GLM routine in R uses Fisher scoring to find maxima. Even when the penalization parameter was set to zero, i.e. unconstrained regression, these two methods can yield rather different results when the likelihood surface is nearly flat. However, even in such cases, the overall model predictions and diagnostics were indistinguishable.

As expected, the estimated filters resembled delta functions for low values of σ, see Figure III.4. In general, the GLMs captured both predictable structure and variability from the input data. For near deterministic spike trains, the variation is negligible, and as such the interpretation and quality of the fit becomes highly sensitive to the choice of basis functions. Other commonly used basis functions, such as splines, which impose smoothness of the influence of past spikes, are likely to lead to poorly fit models and incorrect interpretations, when the spiking process is exactly or nearly deterministic. Although we focus here on simulations, we expect to encounter similar issues when applying the GLM (III.5) to analyze regular spike train activity recorded from a real neuron.

In contrast to low noise regimes, spike trains simulated with a strong, noisy input signal produced spike patterns consistent with a homogeneous Poisson process, cf. Figure III.1. In the corresponding filters for these highly stochastic spike trains, any structure besides a refractory period following an observed spike vanished. For Tonic Spiking there was no positive modulation in very noisy regimes, whereas for Tonic Bursting the intra-burst interval at short delay was captured, while the inter-burst interval at longer delays was missing. This loss of structure was also evident from the relative deviance in Figure III.5 which approached 1 as σ increased.

Our analyses suggest that multiple goodness-of-fit diagnostics are necessary to determine the extent to which statistical models capture the predictable structure and variability of neural spiking. For very low noise regimes, measures that relate primarily to variability, such as the KS statistic, may indicate lack of fit, even for models that capture the predictable structure of the data well. When the random component is ill-fitting, the model does not generalize well to data of similar- but slightly mismatching- structure, thus resulting in poor values of the KS measure. In general, goodness-of-fit measures should account for how well a model captures both the variability (eg. KS test) and predictable structure (e.g. relative deviance) in the data.

Analyzing Tonic Spiking and Tonic Bursting neurons, we found that the GLM captures both types quite well, although the KS statistic remained significant for Tonic Bursting. This was despite the fact that the KS statistics for both types were of equal magnitude, and driven by the higher number of observed spikes for the Tonic Bursting neuron compared to the Tonic Spiking neuron. However, by interpreting the estimated filters, we found that both exhibited a clear structure inherent to the neuron type with features that are intuitively understandable.

There are a number of ways we might extend the models discussed here to better capture structure and variability in the data or to allow for more interpretable parameter estimates. A straightforward model restriction would be to replace the multiplicatively separable effects of multiple previous spikes in the GLM with a renewal model structure,

$$\log\Big(\lambda(t\big|H_t)\Big) = \beta_0 + g(s^*, t), \tag{III.9}$$

where $g(s^*, t)$ is a function of the previous observed spike-time,

$$s^* = \max_j\{s_j|s_j < t\},$$

and time t. With this formulation, estimating a filter $\mathcal{F}$ for (III.9) of a Tonic Spiking neuron would not produce three peaks as in Figures III.2 and III.4, but only a single

peak around the mean ISI at 26.6 ms, since it would only take into account the time since the previous spike. Besides only a single peak for Tonic Spiking, the estimated filter for (III.9) is expected to display similar behavior with regard to convergence and dependence, as the filters for Tonic Spiking examined in this paper. We found that a history dependent filter for Tonic Bursting does not repeat itself, for filters accounting for only 100 ms of past spiking history. However, an estimated filter for (III.9) would not integrate the past history, as was the case in Figure III.3e. This might lead to a more intuitive filter that would display peaks at both bursting and inter-bursting ISIs.

It was clear from the analysis of Tonic Bursting that the estimated filter should be interpreted as a whole and not at individual time points. For the Tonic Bursting neuron, the ISI histogram was bi-modal due to the two time scales involved. This was well captured by the model, as evident from the estimated intensity (Figure III.3b), however the separation of the timescales were not clear from the filter (Figure III.3e). A possible way to extend the model to explicitly account for this separation is to include a latent state that determines when the neuron is bursting. This could be formulated as,

$$\log\Big(\lambda(t|X_t, H_t)\Big) = \beta_0 + \mathbf{1}_{\{X_t=1\}}\Big[\log\Big(\mu(t|H_t)\Big)\Big] + \mathbf{1}_{\{X_t=0\}}\Big[\log\Big(\gamma(t|H_t)\Big)\Big], \quad \text{(III.10)}$$

where

$$X_t = \begin{cases} 1 & \text{neuron is bursting at time } t \\ 0 & \text{else} \end{cases}$$

and $\mu(\cdot|\cdot), \gamma(\cdot|\cdot)$ are individual intensities describing intra-burst and inter-burst times, respectively. A model such as (III.10) could potentially capture both intra-bursting and inter-bursting better than the models considered here, due to the explicit modeling of the dual timescales present. Estimation of the state X_t could also possibly link parameters of such a model to the parameters of the Izhikevich model that controls bursting behavior. The goal in this presentation was to examine a commonly used class of GLMs and analyze their performance in relation to simulated Izhikevich neurons. As such, we leave an analysis of the described model extensions to future work.

In conclusion, the results of this paper suggest that for the commonly used GLM for spike train data with multiplicatively separable history dependence, there is a range of input noise values for the Izhikevich neurons in which the GLM optimally captures both predictable structure and variability. This range depends on which properties are intrinsic to the spike train: intra-burst intervals are captured even at high noise levels, while regular spiking and inter-burst spiking intervals are not captured at high noise levels. At low noise levels, the basis functions implemented here are delta functions, which can predict data structure, but do not generalize well to data from slightly altered generative models. Thus, the choice of basis functions becomes crucial to the credibility of the model for spike trains with little or no variation. However, even though an optimal range is evident from the results in this paper, extensions to the model formulation (III.5), such as (III.9) and (III.10), could possibly lead to improved model descriptions of specific neuron properties, such as bursting and thus extend the optimal range. Expanding the optimal working range of the GLMs would be of interest, for instance in order to classify neurons according to both type and noise level.

5 Acknowledgements

JØ is a participant of the Dynamical Systems Interdisciplinary Network, University of Copenhagen. MAK was partially supported by the NSF CAREER Grant No. NSF-DMS 1451384. UTE was partially supported by a grant from the Simons Foundation (#320136).

MANUSCRIPT IV

A state space model for bursting neurons

Jacob Østergaard, Uri Eden and Susanne Ditlevsen

Working paper.

1 Introduction

Bursting is a complex behavior observed in neural spike dynamics that is characterized by a cluster of rapid successive spikes followed by a longer period of quiescence. It is well known that some classical neuron models such as the FitzHugh-Nagumo model (FitzHugh, 1961) and the Morris-Lecar model (Morris and Lecar, 1981) among others, cannot produce this behavior (Izhikevich, 2004). The duality of this behavior requires a model that can account for dual time scales to handle a fast and slow subsystem such as Izhikevich (2004). However, while the Izhikevich model is capable of producing bursting behavior in simulations, it is cumbersome to fit model parameters to observed data. In contrast, a statistical model class such as the Generalized Linear Model (GLM) is capable of capturing the features of a variety of spike train behaviors as shown in Østergaard, Kramer, and Eden (2018). The flexibility of the GLM is illustrated by the broad applications of this model class (Kass, Eden, and Brown, 2014; McCullagh and Nelder, 1989). However, the interpretation of a GLM is not always intuitive and in Østergaard, Kramer, and Eden (2018) this was indeed the case for a bursting neuron. Although the model was capable of capturing bursting, evident from the estimated intensity, the model itself had a more complex interpretation. In this paper, we show how the GLM can be extended to a State Space GLM (SSGLM) to explicitly account for the dual behavior observed for a bursting neuron. We demonstrate how the model can be fitted to an observed spike train by utilizing a marginalized particle filter to simultaneously decode the state of the neuron (bursting/resting) and estimate history dependent kernels that modulate the baseline firing rate, dependent on the behavior.

The paper is structured as follows. In Section 2 we describe the extended, history dependent and multiplicatively separable GLM, where the state space model formulation includes a latent state controlling bursting. In Section 3 we introduce a marginalized particle filter adapted for point process observations, which is used to simultaneously estimate parameters and decode the latent state process. In Section 4 we present some statistical features of a noisy, bursting Izhikevich neuron and how

these are adapted to the state space GLM. Section 5 presents a simulation study and an analysis of a bursting Izhikevich neuron. Finally Section 6 discuss and conclude the paper.

2 Model description

2.1 Point Process GLM

Let $N_t, t \geq 0$ denote an (arbitrary) counting process of neuron spikes, i.e., N_t denotes the number of spikes fired in the interval $[0, t)$ where the firing rate is given by a, possibly stochastic, time dependent process $\lambda(t|H_t, \theta)$, where H_t denotes the history of past spiking and θ the model parameter. The intensity process is defined by

$$\lambda(t|H_t, \theta) = \lim_{\Delta t \to 0} \frac{P\big(\text{spike in } (t, t+\Delta t] \big| H_t, \theta\big)}{\Delta t}. \tag{IV.1}$$

Let $Y_t = \Delta N_t = N_{t+\Delta t} - N_t$ denote the number of the jumps of N_t in $(t, t+\Delta t]$. For sufficiently small Δt, $P(Y_t > 0) \approx \lambda(t|H_t, \theta)\Delta t$ and $P(Y_t > 1) \approx 0$, hence the binomial distribution can describe the spike distribution in each time bin of size Δt. However, since the binomial and Poisson distribution converge in the limit of $\Delta t \to 0$, the Poisson distribution is another valid choice to model the probability of spiking in each time bin. Since the Poisson distribution leads to a more intuitive interpretation of the model parameters due to the canonical link function, this choice of distribution is preferable to the binomial.

Given observations $\{y_k\}_{k=1}^K$, where $y_k \in \{0, 1\}$, of Y_t, for $t \in [0, T]$, at a given sectioning of the observation interval, $0 < t_1 < \cdots < t_K = T$, where $t_k - t_{k-1} = \Delta t$, the firing rate can be modeled as

$$f\big(\lambda(t_k|H_{t_k}, \theta)\big) = \beta_0 + \sum_{i \in \mathbb{N}} h(k-i) y_{k-i}, \tag{IV.2}$$

where f is a possibly nonlinear link function, $f^{-1}(\beta_0)$ specifies the baseline firing rate and the kernel $h(k-i)$ characterizes the influence of past spikes. In the following we shall refer to (IV.2) as the Point-Process GLM (PPGLM). The Poisson distribution choice leads to $f(\cdot) = \log(\cdot)$, the canonical link function for the Poisson GLM. Hence, $\exp \beta_0$ becomes the baseline firing rate and $h(k-i)$ controls the influence of the past observations, y_{k-i}, on the modulation of the baseline firing rate. The history dependency in (IV.2) implies that Y_t is a non-homogeneous Poisson process which is necessary to include effects such as refractoriness in the behavior of Y_t. The simplest kernel $h(k-i)$ is $h(k-i) = \beta_i \in \mathbb{R}, i = 1, 2, \ldots$. However, this choice is not feasible for a model where the length of the memory is substantial, since a parameter is needed at each past timepoint. In the case of long memory, it is advantageous to reduce the dimension of (IV.2) by modeling $h(k-i)$ with basis functions. Significantly reducing the number of parameters can then improve the estimation of the model. Splines are a common choice of basis functions that impose a continuity assumption on the effect of the past history.

2.2 State space model for bimodal ISI distributions

Although the PPGLM model is very flexible despite the simple construction (IV.2) and can capture multiple neuron behaviors as shown in Østergaard, Kramer, and Eden (2018), it does not seem to be able to account for the dual behavior of bursting neurons. We therefore propose an extended version of (IV.2) where the model is augmented with a latent state, controlling the current behavior by switching between a "burst" mode and a "resting" mode. If the neuron behavior can be split into two behavioral states, defined by a latent process X_t, we can reformulate the model as

$$\begin{aligned} f\big(\lambda(t_k|H_{t_k},\theta,X_{t_k})\big) =& \mathbb{1}\{\text{rest}|\ X_{t_k}\}\Big(\beta_0^r + \sum_{i=1}^{l} h^r(k-i)y_{k-i}\Big) \\ &+ \mathbb{1}\{\text{burst}|\ X_{t_k}\}\Big(\beta_0^b + \sum_{i=1}^{l} h^b(k-i)y_{k-i}\Big), \end{aligned} \tag{IV.3}$$

where the superscripts r and b refer to *resting* and *bursting*, respectively, and $\mathbb{1}\{\cdot\}$ denotes the indicator function. The summation to l implies that the model only includes the past history up to this lag, and past history beyond this point, i.e. y_{k-i}, $i > l$, does not influence the intensity at time k. We will refer to model (IV.3) as a State Space GLM (SSGLM). The latent process determines whether the neuron is bursting or resting such that the model includes separate kernels for these modes and thus explicitly reveal the structure of each mode. We opt for a spline representation of $h(k-i)$ to keep the number of parameters tractable, while enabling a flexible form. For a given set of p basis functions (splines), stored as columns in $l \times p$ matrix S, the corresponding parameters are given as a vector of weights $(\beta_1, \ldots, \beta_p)'$, such that each kernel (superscript suppressed) is represented as

$$h(k-i) = \sum_{j=1}^{p} S_{ij}\beta_j \quad \text{for } i = 1, \ldots, l, \tag{IV.4}$$

where S_{ij} denotes the j'th element of row i. Using superscripts r, b, the total number of parameters of the model is then $p^r + p^b + 2$, where the 2 extra dimensions represents the baselines. Note that if $h^r(k-i) = 0$, i.e., $p^r = 0$, then the model assumes that the neuron only spikes according to some baseline firing rate, possibly also 0, when resting.

In this paper we use cardinal splines to represent the kernels h^r, h^b in (IV.3) and (IV.4). As the kernel superscripts r and b refer to resting and bursting kernels, respectively, we use the same notation for the spline basis S^r, S^b and parameters β_j^r, β_j^b. In order to pick a suitable set of basis functions for each kernel, we used a histogram of the interspike intervals for a specific neuron (Figure IV.3) to set the knot locations. Other methods exist to estimate knot locations, such as BARS (see Dimatteo, Genovese, and Kass, 2001), but here we focus primarily on decoding the latent process controlling bursting as well as the kernel weights. Thus, more sophisticated strategies of selecting knots for this model is left for future research. The simple representation of the kernels in this manuscript ensures a simple, yet flexible, model structure capable of capturing the bimodal ISI distribution.

To construct a latent process X_k that determines the neuron behavior, we define $X_k \in \mathbb{Z}$ as "the length of a current bursting or resting period at time t_k", where the

sign of X_k determines whether the neuron is resting ($X_k \leq 0$) or bursting ($X_k > 0$), such that

$$\mathbb{1}\{\text{resting}\} = \mathbb{1}\{X_k \leq 0\}$$
$$\mathbb{1}\{\text{bursting}\} = \mathbb{1}\{X_k > 0\}.$$

This definition of X_k allows for a time dependent structure in the mode switch, which is necessary for a realistic model. As the length of a burst/rest period increases, the probability of switching should increase. Thus,

$$\begin{aligned} P\big(X_k = x-1 \big| X_{k-1} = x \leq 0\big) &= g^r(x) = 1 - P\big(X_k = 1 \big| X_{k-1} = x \leq 0\big) \\ P\big(X_k = x+1 \big| X_{k-1} = x > 0\big) &= g^b(x) = 1 - P\big(X_k = 0 \big| X_{k-1} = x > 0\big), \end{aligned} \tag{IV.5}$$

for some, possibly nonlinear, decreasing functions g^r, g^b. Had we opted for a binary state where $X_k = 0, 1$ denotes resting and bursting, respectively, then realistically X_k and X_s for $s < k-1$ should not be assumed independent. In that case it would be necessary to include a higher order Markov structure in the latent process, thus complicating estimation. However, by including the time memory in the definition of X_k, we can assume that X_k is a Markov process. In both cases of (IV.5), X_k the resetting is defined such that X_k either increase/decrease or switch to the other mode. Hence, the probability of switching is $1 - g^r(x)$ and $1 - g^b(x)$, respectively. At each timepoint, given X_k, we define the probability of spiking for the observed process Y_k as

$$\begin{aligned} P(Y_k = 1 | X_k \leq 0, H_k) = p^r(H_k) \approx \lambda(t_k | H_k, \theta, X_k \leq 0)\Delta t \\ P(Y_k = 1 | X_k > 0, H_k) = p^b(H_k) \approx \lambda(t_k | H_k, \theta, X_k > 0)\Delta t \end{aligned}, \tag{IV.6}$$

where the two probabilities p^r and p^b treat the history of Y_k differently, as measured by the two distinct kernels in (IV.3). If we let $p^r(H_k) = 0$, this implies that the neuron cannot spike when resting. The dependence graph of the model is displayed in Figure IV.1.

3 Filtering and estimation of the SSGLM

For a general state space model with observations y_k dependent on a latent state x_k, the problem of decoding the current state x_k given observations has applications in many fields (Cappé, Moulines, and Ryden, 2005). Given a set of observations $y_{1:s} = (y_1, \ldots, y_s)$ what is the best estimate $\hat{x}_k$ of x_k, for $k = 1, \ldots, K$? Depending on the phrasing of the problem, this estimate is known as either *filtering* ($s = k$), *smoothing* ($s = K$), *fixed lag smoothing* ($s = k + n$ for some fixed $n \in \mathbb{N}$) or *prediction* ($s = k - 1$). When $s = k$ and both x_k and y_k have linear dynamics and Gaussian errors, the Kalman filter (Kalman, 1960) provides an exact solution for the posterior density of x_k conditional on the observations. Let θ denote the model parameter. Decomposing the posterior density of the state process as

$$p_\theta(x_k | y_{1:k}) = \frac{p_\theta(y_k |\, x_k) p_\theta(x_k |\, y_{1:k-1})}{p_\theta(y_k | y_{1:k-1})}, \tag{IV.7}$$

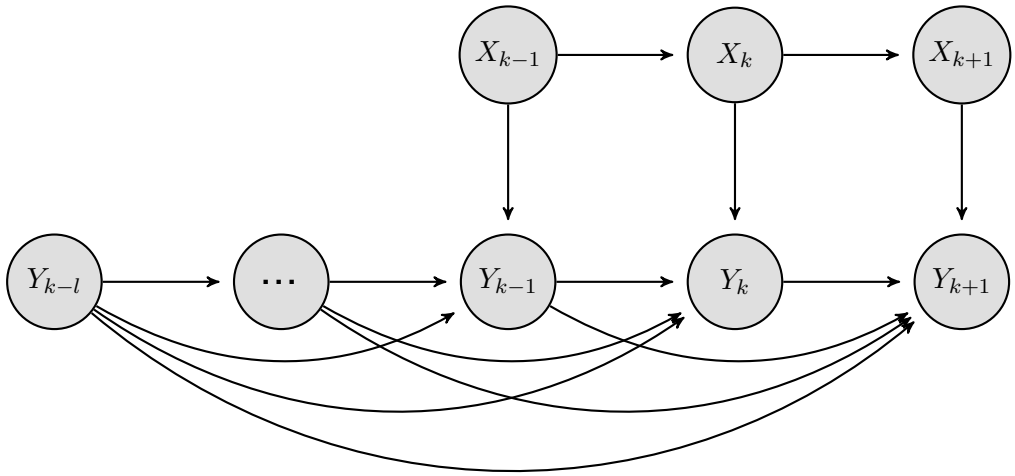

FIGURE IV.1: State Space Model of neurons with multimodal ISI distributions and history dependent intensities. The process $Y_k = Y_{t_k}$ depends on past values of itself as well as the current state X_k. The latent states control the multiple behaviors that cause the multimodal ISI distribution.

with

$$p_\theta(x_k|\ y_{1:k-1}) = \int p_\theta(x_k|x_{k-1}) p_\theta(x_{k-1}|y_{1:k-1}) dx_{k-1}, \tag{IV.8}$$

then for a Gaussian prior $p_\theta(x_k|x_{k-1})$ and Gaussian observation density $p_\theta(y_k|\ x_k)$, the posterior is Gaussian and there exist an analytical expression for (IV.7) which is the Kalman filter. However, when the distribution of either x_k, y_k are nonlinear and/or non-Gaussian, the Kalman filter is not directly applicable. If the errors are Gaussian, or can be adequately approximated as Gaussian, then by linearizing the dynamics of the model, the Kalman filter can be applied. This approach is termed the extended Kalman filter, but contrary to the standard Kalman filter, it does not imply an exact posterior distribution, due to the approximation. Another approach is to use a particle filter (Cappé, Moulines, and Ryden, 2005) to approximate the posterior density (IV.7). The standard particle filter samples particles from a proposal (or importance) distribution $q(x_k)$ and weigh these according to the observation likelihood, thus providing an approximation of the posterior density (IV.7).

3.1 Marginal Particle Filter for Point Process observations

The particle filter is feasible, although computationally much more expensive than the (extended) Kalman filter. In order to obtain an adequate performance when the dimension of the latent state increases, the number of particles must also be increased, thus improving the numerical performance of the filtering. While this might not be a problem theoretically, it can lead to problems in the practical evaluation. When θ in (IV.7) and (IV.8) is unknown, then besides decoding the latent process, parameter estimation is necessary. A strategy for this is to include the parameter θ as a latent process, thus augmenting the model to simultaneously estimate x_k and θ. This further increases the dimension of the number of latent states, and hence a pure particle filter might be impractical due to the number of particles required for

satisfactory performance. A way to cope with this issue is to marginalize the particle filter, which means that the latent state is split into two parts. One that can be linearized and thus a version of the Kalman filter can be applied, and another where a linearization is not feasible and thus a particle filter must be applied.

For model (IV.3) the parameter consists of the baselines β_0^r, β_0^b and the weights for the kernels (IV.4)

$$\theta = (\beta_0^r, \beta_1^r, \ldots, \beta_{p^r}^r, \beta_0^b, \beta_1^b, \ldots, \beta_{p^b}^b)' \in \mathbb{R}^{p^r+p^b+2}. \tag{IV.9}$$

The augmented state is thus,

$$\xi_k = (X_k, \theta_k)', \text{ for } k = 1, \ldots, K, \tag{IV.10}$$

where we have implicitly assumed a temporal evolution of the parameter estimate θ_k. Assuming that we can partition ξ_k as

$$\xi_k = (\xi_k^{\text{non}}, \xi_k^{\text{lin}})', \tag{IV.11}$$

where the superscripts 'non' and 'lin' refer to the non-linear and linear parts of ξ_k, respectively, we can exploit that ξ_k has a linear subspace by applying a marginalized particle filter, thus keeping the dimensionality under control. This solution is, of course, mostly tractable when the majority of the state space can be linearized, but this is precisely the case if we assign a Gaussian prior and a random walk evolution for the parameter θ.

Computationally, we split the joint posterior density of ξ_k,

$$p(\xi_k^{\text{non}}, \xi_k^{\text{lin}} |\, y_{1:k}) = p(\xi_k^{\text{lin}} |\, \xi_k^{\text{non}}, y_{1:k}) p(\xi_k^{\text{non}} | y_{1:k}). \tag{IV.12}$$

From a sample of particles $\xi_k^{\text{non},m}, m = 1, \ldots, M$ we apply a version of the Kalman filter to $p(\xi_k^{\text{lin},m} |\, \xi_k^{\text{non},m}, y_{1:k})$ for each particle. From the weights

$$w_k^m = p(y_k |\, \xi_k^{\text{non},m}, \xi_k^{\text{lin},m}, y_{1:k-1}), m = 1, \ldots, M, \tag{IV.13}$$

we obtain the estimate

$$p(\, \xi_k^{\text{lin}} |\, \xi_k^{\text{non}}, y_{1:k}) \approx \sum_{m=1}^{M} w_k^m p(\, \xi_k^{\text{lin},m} |\, \xi_k^{\text{non},m}, y_{1:k}) \tag{IV.14}$$

as a weighted mixture of the linear posteriors conditional on each particle.

3.2 Stochastic State Point Process Filter

Since the observed process Y_k is modeled as a point process, then in order to exploit the partial linearization of ξ_k in (IV.11), it is necessary to define an analogue version of the Kalman filter for point processes. Consider the PPGLM (IV.2) with corresponding parameter $\theta = (\beta_0, \beta_1, \ldots, \beta_p)' \in \mathbb{R}^{p+1}$. We then redefine the parameter θ as an evolving process

$$\theta_{k+1} = F_k \theta_k + \varepsilon_k, \text{ for } k = 1, \ldots, K \tag{IV.15}$$

where F_k is a matrix chosen specifically to reflect the evolution of θ. If no correlation structure between the elements of θ is assumed, choosing $F_k = I_{p+1}$, where I_d denote the d-dimensional identity matrix, is a valid choice to initially omit any interdependence in θ. The ε_k is chosen as a multivariate zero-mean white noise process with (known) covariance matrix Q which is also chosen initially. As in the case of F_k, choosing $Q = \sigma^2 I_{p+1}$ for some $\sigma > 0$, assumes that the elements in θ does not depend on each other in the prior distribution. Note that choosing these settings for the prior, does not necessarily lead to independent elements in the posterior distribution of θ, which is the prior updated with information from the observations.

Let $\theta_{k|k-1}$ and $W_{k|k-1}$ denote the mean and covariance of the prior distribution at time k. Likewise, let $\theta_{k|k}$, $W_{k|k}$ denote the mean and variance of the posterior distribution at time k.

Using this notation for the prior and posterier estimates of θ, Eden et al. (2004) derives the stochastic state point process filter

$$\theta_{k|k-1} = F_k \theta_{k-1|k-1} \tag{IV.16a}$$

$$W_{k|k-1} = F_k W_{k-1|k-1} F_k' + Q_k \tag{IV.16b}$$

$$W_{k|k}^{-1} = W_{k|k-1}^{-1} + \left[\left(\frac{\partial \log \lambda}{\partial \theta_k} \right)' [\lambda \Delta t] \left(\frac{\partial \log \lambda}{\partial \theta_k} \right) - (y_k - \lambda \Delta t) \frac{\partial^2 \log \lambda}{\partial \theta_k \partial \theta_k'} \right]_{\theta_{k|k-1}} \tag{IV.16c}$$

$$\theta_{k|k} = \theta_{k|k-1} + W_{k|k} \left[\left(\frac{\partial \log \lambda}{\partial \theta_k} \right)' (y_k - \lambda \Delta t) \right]_{\theta_{k|k-1}}, \tag{IV.16d}$$

where λ is shorthand notation for $\lambda(t_k|\theta_k)$ in (IV.2). Equations (IV.16a) and (IV.16b) show the (time) prediction step which is given by the prior distribution defined by (IV.15). Equations (IV.16c) and (IV.16d) define the (observation) update step, defined by the posterior, which is assumed Gaussian. Starting from initial values $\theta_{0|0}, W_{0|0}$ the algorithm iterates through the observations y_k, for $k = 1, \ldots, K$ alternating between predicting and updating the parameter estimate θ_k. Thus, the SSPPF can be viewed as a version of the extended Kalman filter for point processes where the observation density is approximated with a Gaussian.

3.3 The Marginal Particle Filter for Point Process observations

We use the point process version of the extended Kalman filter for the linearized part of ξ_k and a particle filter for the non-linear part of ξ_k and term this version of a marginalized particle filter the *Marginal Particle Filter for Point Process observations* (MPFPP). Let $\theta_{k|k-1}^m, \theta_{k|k}^m, W_{k|k-1}^m, W_{k|k}^m, F_k^m, Q_k^m$ and λ^m denote the variables in equations (IV.16a)-(IV.16d) for the m'th particle $\xi_k^{\text{non},m} = x_k^m$. For a general importance density $q(x_{1:k})$, updating importance weights is given as

$$w_k^m = w_{k-1}^m \frac{p(y_k|\, x_k^m, y_{1:k-1}) p(x_k^m|\, x_{k-1}^m)}{q(x_{1:k}^m)}, \tag{IV.17}$$

but when using the bootstrap particle filter, the proposal distribution for the particle samples is exactly the transition density, therefore (IV.17) simplifies to

$$w_k^m = w_{k-1}^m p(y_k|\, x_k^m, y_{1:k-1}). \tag{IV.18}$$

The algorithm, adapted from Schon, Gustafsson, and Nordlund (2005), is then as follows.

Initialize $\theta_{0|0}^m = \theta_0$, $W_{0|0}^m = W_0$ for $m = 1, \ldots, M$ and sample particles $\{x_0^m\}_{m=1}^M$ from an initial distribution $x_0^m \sim p(x_0)$, set all weights $w_0^m = M^{-1}$. Then for $k = 1, \ldots, K$ and each $m = 1, \ldots, M$ proceed with

1. Sample $x_k^m \sim p(x_k|x_{k-1}^m, y_{k-1})$.
2. Predict $\theta_{k|k-1}^m, W_{k|k-1}^m$ by (IV.16a) and (IV.16b).
3. Update importance weights w_k^m according to (IV.18) and normalize $\tilde{w}_k^m = \frac{w_k^m}{\sum_m w_k^m}$.
4. Resample the full particles $\xi_{1:k}^m = (x_{1:k}^m, \theta_{1:k|0:k-1}^m)$ according to the normalized weights $\tilde{w}_k^m$ and reset weights $w_k^m = M^{-1}$.
5. Update the linear posteriors $W_{k|k}^m$ and $\theta_{k|k}^m$ using (IV.16c) and (IV.16d).
6. Set $k = k + 1$ and go to step 1.

The resampling step is optional and can be omitted. It ensures that the weights w_k^m are not degenerate, but may not be necessary at each step. A commonly used measure for degeneracy is the efficient sample size (Cappé, Moulines, and Ryden, 2005)

$$N_{\text{eff}} = \Big(\sum_{m=1}^{M} (w_k^m)^2 \Big)^{-1}, \tag{IV.19}$$

ranging from 1 (only one effective particle) to M (all particles are equally weighted and effective). Hence, the resampling step can be modified, such that resampling is only performed whenever N_{eff} is below a chosen threshold.

4 Bursting model for Izhikevich neurons

We adapt model (IV.3) to the specific task of modeling Izhikevich neurons. The Izhikevich (2003) model is a well known model capable of replicating multiple types of behavior observed in real neurons.

The model is a simple two-dimensional dynamical system, where the two variables v_t and u_t evolve according to

$$\begin{aligned} \frac{dv_t}{dt} &= 0.04v_t^2 + 5v_t + 140 - u_t + I_t \\ \frac{du_t}{dt} &= a(bv_t - u_t) \\ \text{if } v_t \geq 30 \text{ then } & \begin{matrix} v_{t+} = c \\ u_{t+} = u_t + d, \end{matrix} \end{aligned} \tag{IV.20}$$

for given parameters a, b, c, d and input current $I_t, t \geq 0$ (see Izhikevich, 2003), with the timescale interpreted in ms. A spike is defined as occurring whenever $v_t \geq 30$, hence the resetting mechanism in (IV.20). Here we introduce noise by letting $I_t = I_0 + \sigma\varepsilon_t$, where ε_t is a standard 1-dimensional white noise process. The two variables v_t and u_t allow for the dual behavior of bursting neurons and as such it is an easy model to generate a realistic sample from a bursting neuron.

4.1 Features of a noisy bursting Izhikevich neuron

In order to define realistic transition functions g^r, g^b in (IV.5), we use (IV.20) to generate 20 s of observations, with a 500 ms burn-in period, from a bursting neuron and characterize the features related to bursting, such as the maximal burst length and the speed of decay for the probability of continued bursting. To reduce the discretization error when simulating the model with stochastic noise in I_t, the time discretization was set to 10^{-3} ms for simulation, but with a subsequent subsampling we obtained a discretized time step of $\Delta t = 0.1$ ms. The parameters were set to $a = 0.02, b = 0.2, c = -50, d = 2$ and $I_0 = 10$, taken from Izhikevich (2004). The noise amplitude was set at $\sigma = 100$ to create a noisy spike train and the initial conditions were $v_0 = -70, u_0 = bv_0$.

Figure IV.2 presents the first 5000 ms of a simulated spike train from (IV.20) using an Euler-Maruyama scheme.

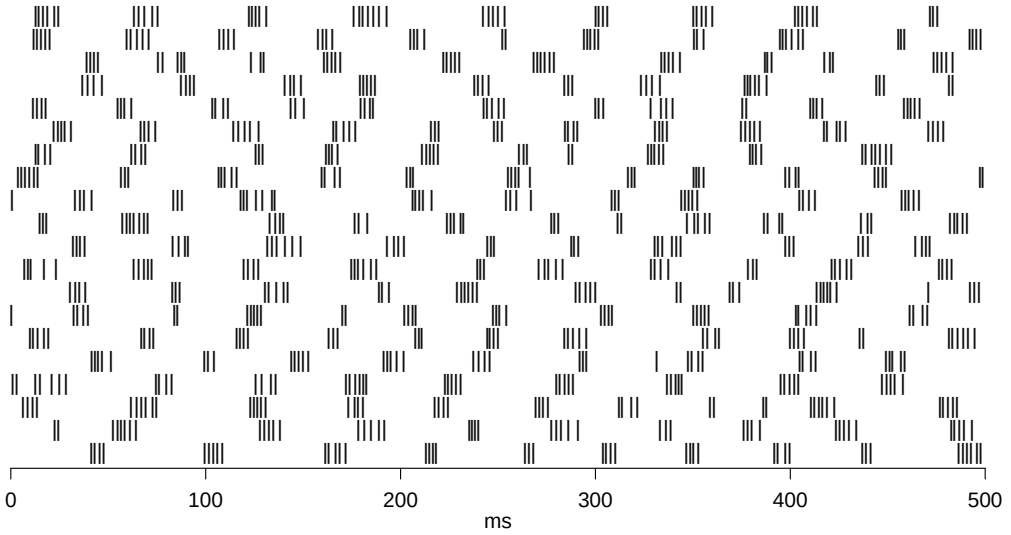

FIGURE IV.2: First 5000ms (50 000 observations) from a bursting Izhikevich neuron. To be read left to right from the top, like lines in a book. The bursts are clearly identified as the spike clusters with varying ISIs between and within bursts. Note that some bursts consist only of one spike, due to the stochastic input.

Figure IV.3 shows the histogram of ISIs of the entire spike train, partly shown in Figure IV.2. The bi-modal structure indicates the dual time scales involved, with the left peak representing the bursting periods of short ISIs and fast spiking and the right peak representing the resting period in between bursts. Figure IV.4a presents a histogram of the length of the bursts from the simulated neuron. It is evident that most bursts are around 4-15 ms in length, with a few longer/shorter bursts. The distribution of resting intervals, between bursts, is shown in Figure IV.4b and reveals a

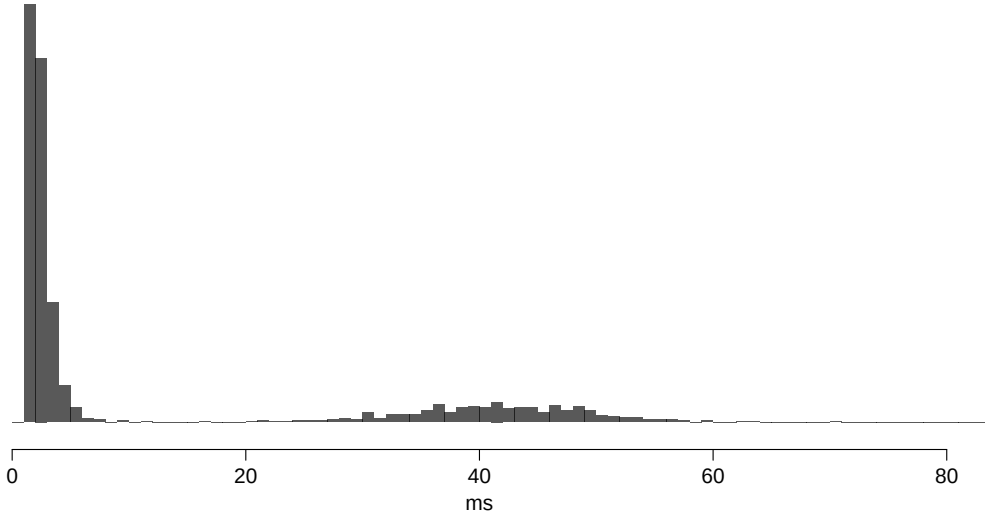

FIGURE IV.3: Interspike-interval histogram for the bursting neuron, partly shown in Figure IV.2.

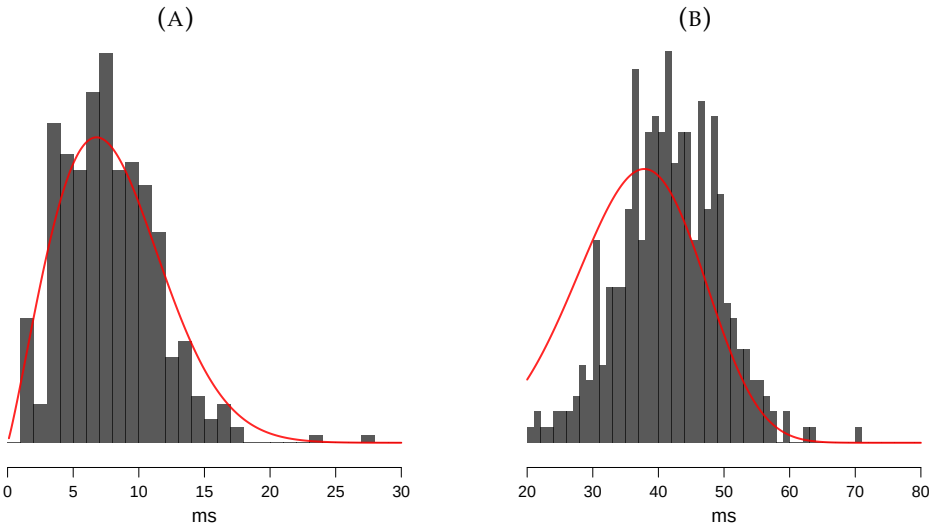

FIGURE IV.4: Descriptive histograms of the bursting neuron, partly shown in Fig. IV.2. (A) Distribution of the length of bursting periods. (B) Distribution of the length of resting periods. Both empirical distributions are approximated with parametrized functions (red curves), see text.

slightly skewed distribution with the mode around 40 ms. The empirical distribution of burst/rest lengths where derived from the spike train by setting a maximum ISI of 20 ms within a burst. By splitting the observed ISIs, it is possible to extract the length of bursts and resting periods respectively. From the ISI histogram in Figure IV.3 20 ms does not seem unreasonable when splitting between bursting/resting to identify the bursting/resting length distributions.

The transition probabilities (IV.5) were modeled using a log-logistic function

$$g(x, \alpha, \kappa) = \frac{x^\kappa}{\alpha^\kappa + x^\kappa}, \text{ for } x = 1, 2, \ldots, \tag{IV.21}$$

fitted to empirical values of a bursting Izhikevich neuron. Parameter values $\alpha > 0$ and $\kappa < 0$ implies, that the function is decreasing as a function of x, which denotes the length of the current burst or resting period. The log-logistic function is suitable to capture some of the skewness apparent in the distributions in Figure IV.4, apparent by the fitted densities (IV.25) with parameters

$$\begin{aligned} \alpha^r &= 1480 \\ \kappa^r &= -3.45 \\ \alpha^b &= 2541 \\ \kappa^b &= -1.13 \end{aligned} \tag{IV.22}$$

The parameters of the transition probabilities $g^r(x, \alpha^r, \kappa^r), g^b(x, \alpha^b, \kappa^b)$ were fitted by minimizing the squared difference between the empirical distributions in Figure IV.4 and (IV.25), see Appendix A. The approximating distributions in Figure (IV.4) capture some of the skewness in the empirical distributions.

5 Results

We opted for a model where $h^r(\cdot) = \beta_0^r = 0$, i.e., no spikes can occur when the neuron is resting. This was to facilitate a simpler parameter estimation when spikes only occur during bursting, implying that the h^r kernel is in fact redundant.

For the propagation model, equations (IV.16a)-(IV.16d), we set $F_k = I_p$, where $p = p^b + 1$ is the dimension of the parameter θ consisting of weights for the bursting kernels and a baseline parameter. The covariance of the propagation model was set to $Q = 10^{-7} \cdot \text{diag}(1, \ldots, 1)$. From (IV.3)-(IV.4) and (IV.9) the derivative of $\log \lambda_k$ is given as

$$\frac{\partial \log \lambda_k}{\partial \theta_k} = \mathbb{1}\{X_k > 0\} \begin{pmatrix} 1 \\ \sum_{i=1}^l S_{i1}^b y_{t-i} \\ \vdots \\ \sum_{i=1}^l S_{ip^b}^b y_{t-i} \end{pmatrix} \in \mathbb{R}^p, \tag{IV.23}$$

and from (IV.23) it is clear that

$$\frac{\partial^2 \log \lambda_k}{\partial \theta_k \partial \theta_k'} = 0 \in \mathbb{R}^{d \times d}. \tag{IV.24}$$

The lag length was set to 10 ms ($l = 100$ for $\Delta t = 0.1$ms) and two spline knots were set at 2 ms and 5 ms. An auxiliary boundary knot was set at 0 ms, with fixed weight of -10, to impose a refractory period. Hence,

$$h(k-i) = \sum_{j=1}^{3} S_{ij}\beta_j, \text{ for } i = 1, \ldots, l,$$

where $\beta_1 = -10$. The dimension of θ is thus $p = 2 + 1 = 3$, since β_1 is fixed. The spline weights $\beta_2 = 1.17$ and $\beta_3 = -0.49$ along with $\beta_0 = 0.88$ were tuned[1] to mimic the properties of a bursting Izhikevich neuron. The simulation kernel h^b is shown in Figure IV.5. For the numerical analysis, we fixed the parameters α and κ in the

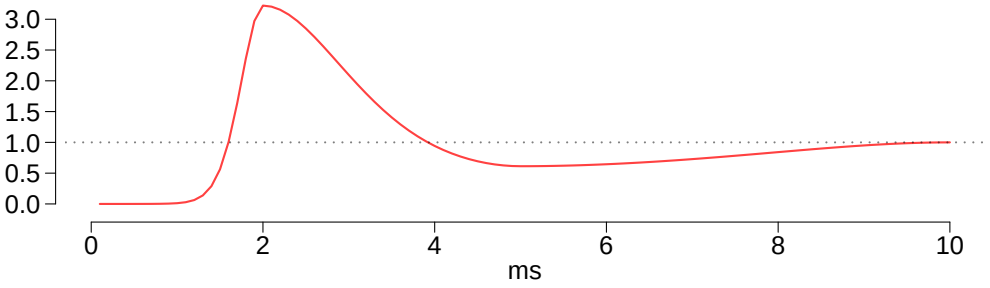

FIGURE IV.5: Bursting kernel used for simulation of a SSGLM neuron, in the exponential domain, i.e., $\exp(h^b)$. The resting kernel h^r and β_0^r was set to 0, thus no spikes occurred while the neuron was resting.

log-logistic functions corresponding to the values in Figure IV.4, thus treating the transition probability parameters as known. This leaves the latent state X_k as the only nonlinear component for the marginalized particle filter. For the particle filter, the number of particles was set to $M = 50$.

5.1 Simulation model

A time period of 20 s was simulated from the SSGLM model, with a time step of $\Delta t = 0.1$ ms. Figure IV.6 compares the resulting ISI distribution (red) with the one in Figure IV.3 from the Izhikevich neuron (gray). The ISI histograms compares well and the features regarding bursting and resting are close. Figure IV.6 shows that the SSGLM neuron (red) have a few large observed ISIs (> 70 ms), but the majority of the distribution compares well to the Izhikevich ISI distribution (gray). Figures IV.7a and IV.7b compares the distribution of the bursting and resting periods, respectively. The distributions of burst lengths (Figure IV.7a) shows that the SSGLM neuron (red) captures the Izhikevich neuron (gray) well. This is also evident from the medians of 7.60 ms (SSGLM) and 7.60 ms (Izhikevich) respectively. Comparing the distributions of rest lengths for the SSGLM neuron (red) and the Izhikevich neuron (gray) they agree to some extend, except for the few very long resting periods above 80 ms. The mode of the Izhikevich resting length distribution appears slightly to the right of the SSGLM resting length distribution. The medians of the rest period distributions are 41.40 (SSGLM) and 41.80 (Izhikevich), respectively, and thus comparable although the tails of the empirical SSGLM resting length distribution are thicker. Overall these

[1] By applying the MPFPP algorithm in an initial exploratory study of an Izhikevich spike train.

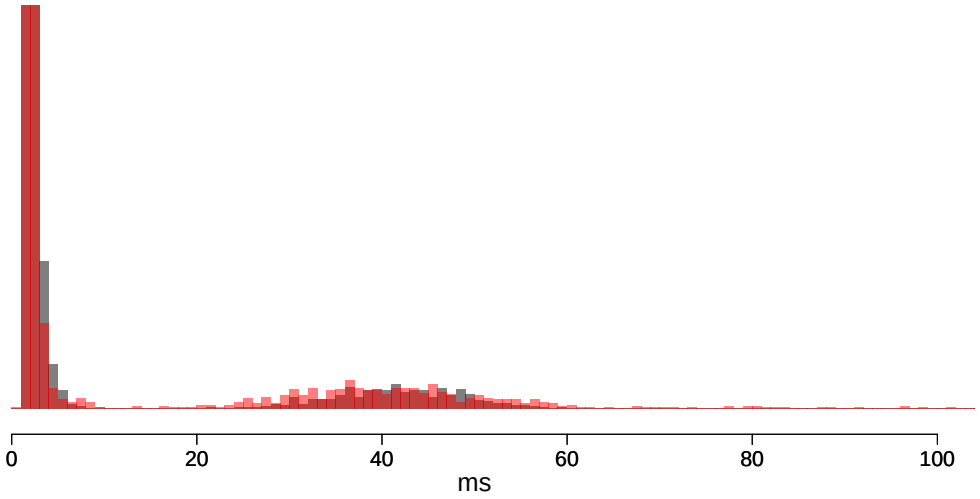

FIGURE IV.6: Comparison of ISI distributions. The histogram in Figure IV.3 of a bursting Izhikevich neuron (gray) is superimposed with a histogram from an SSGLM neuron (red). The bimodal structure is well captured, although the SSGLM neuron exhibit more very long ISIs.

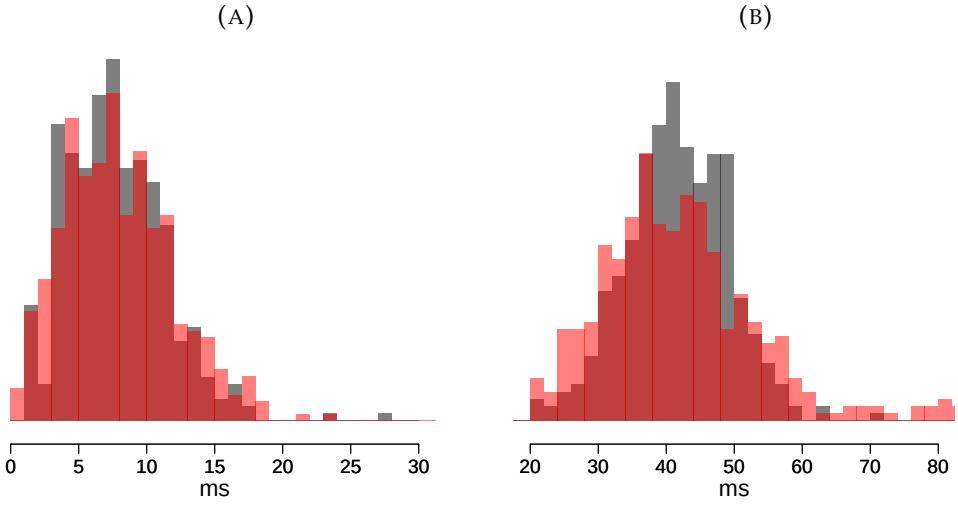

FIGURE IV.7: Comparing simulations of a bursting Izhikevich neuron (gray) and a SSGLM neuron (red). (A) Burst length distributions. (B) Rest length distributions. The Izhikevich histograms correspond to the ones in Figure IV.4, superimposed with the corresponding histograms of the SSGLM simulation, hence observed burst lengths > 80 ms of the SSGLM neuron are not visible in (B).

findings show that the SSGLM neuron is capable of displaying similar qualitative behavior as the bursting Izhikevich neuron.

5.2 Parameter estimation and decoding

We compare the performance of the MPFPP algorithm using observations from a simulated SSGLM neuron where both the true baseline and kernel, as well as the true latent process, are known. As above, 20 s of observations where used for the MPFPP algorithm, but the data was repeated twice to obtain 40 s in total. This was to observe whether the parameter convergence was sufficient at 20 s. The parameter estimates where set as the average of the final 10 s corresponding to the last quarter of the values from the MPFPP algorithm. As mentioned above, β_1 was fixed at -10, whereas the estimated parameters β_0, β_2 and β_3 were initiated at 0 to reflect no prior knowledge on their levels.

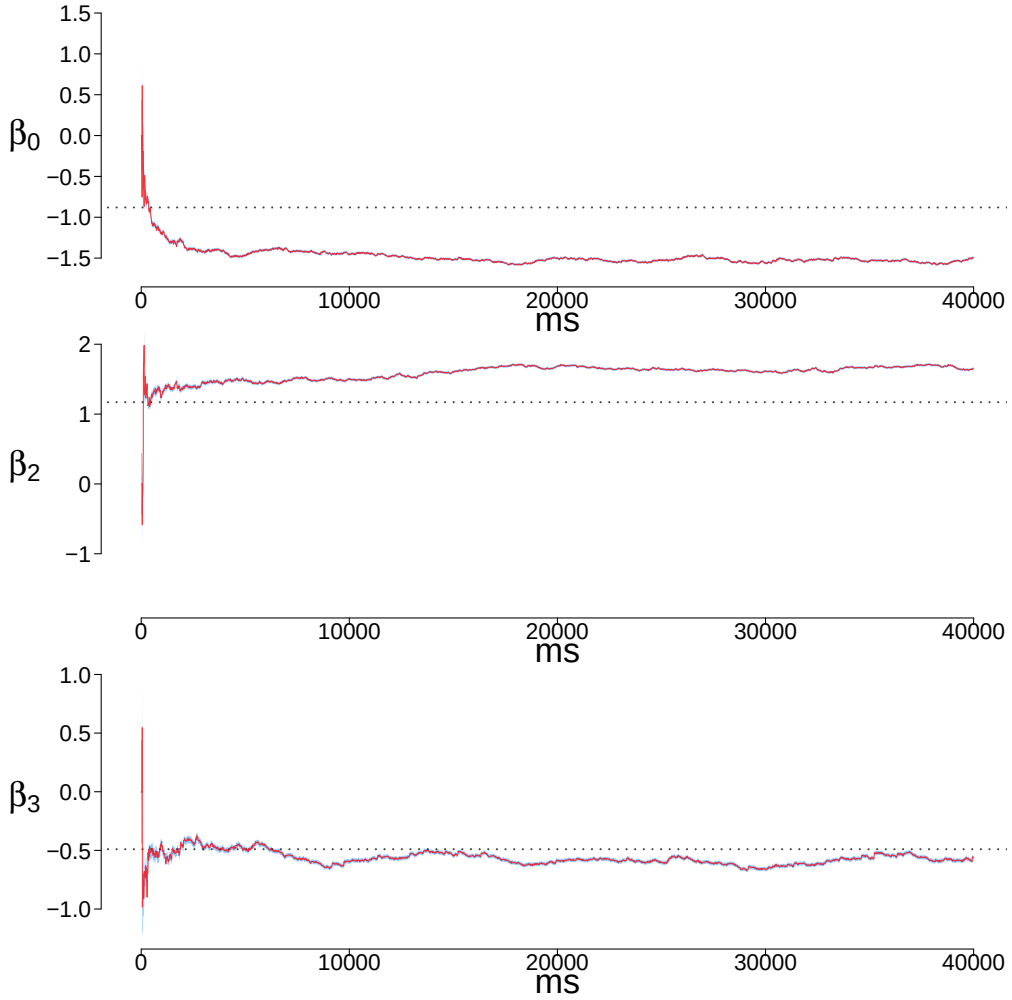

FIGURE IV.8: Convergence of parameters using 20 s of observations repeated once. Parameter estimates in red, true values as horizontal black dashed lines. Convergence seems to be decent already at 20 s, hence the recycling of input data is not necessary. The baseline parameter β_0 is underestimated, but this is partially offset by the overestimation of β_2. The β_3 parameter show less consistent convergence.

Figure IV.8 displays the convergence of parameters $\beta_j, j = 0, 2, 3$ for the total 40 s of data. The baseline parameter β_0 converges to a level below the true value and β_2 converges to a level above the true value. The imprecise convergence of these two parameters partially offset each other as seen in Figure IV.9 below. The β_3 parameter displays less steady convergence as evident from the low dip around both 10 s and 30 s, both times correspond to halfway through the original data, since 30 s compares to 10 s when the data is copied onto itself. However, the parameter estimation of β_3 fluctuates close to the true level, although slightly below. All three parameters seem to reach convergence at 20 s, only β_3 is marginally closer to the true value at the end of the second period of the data, at 40 s. Therefore, recycling the data doesn't appear necessary in this case.

Figure IV.9 shows the estimated kernel (red) in the exponential domain, without the baseline (top) and with the baseline added (bottom). These were based on parameter estimates (true values) $\hat{\beta}_0 = -1.53$ (-0.88), $\hat{\beta}_2 = 1.65$ (1.17) and $\hat{\beta}_3 = -0.58$ (-0.49). Comparing to the true kernel (black) without the baseline addon, the overestimation

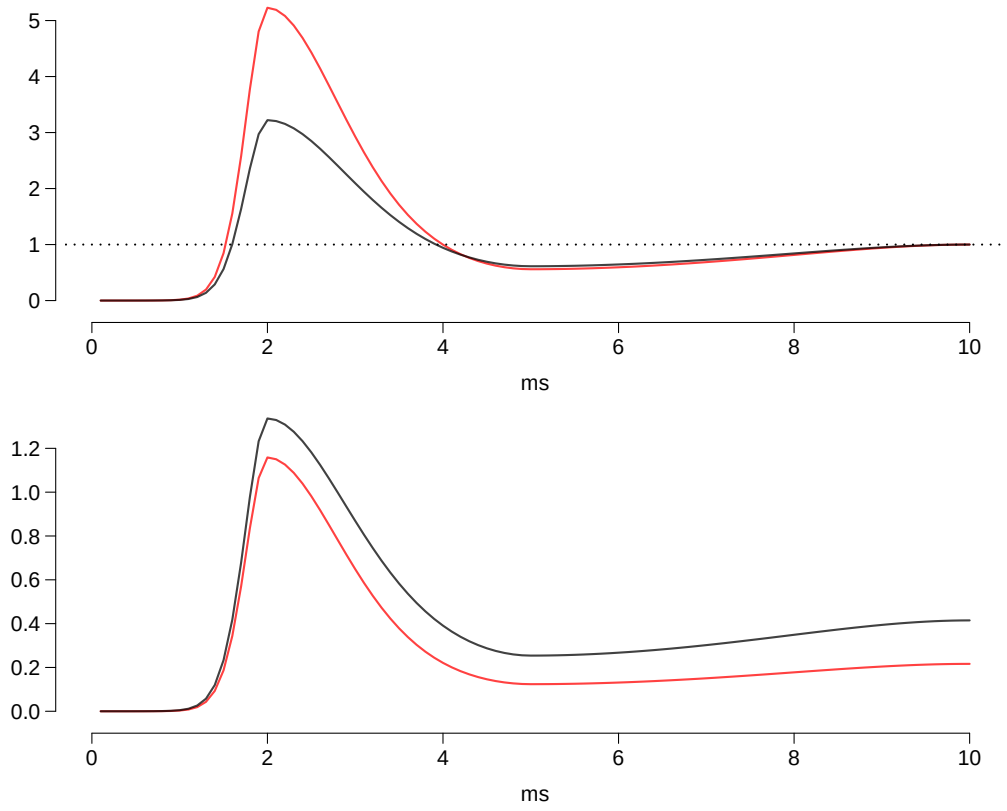

FIGURE IV.9: Estimated kernel (red) in the exponential domain, based on the last 10 s versus simulation kernel (black). Top: estimated kernel without baseline effect. Bottom: estimated kernel with baseline effect.

of β_2 (spline weight at 2 ms) is evident around 2-4 ms, whereas the right tail of the kernel is well approximated due to a relatively precise estimate of β_3 (spline weight at 5 ms). However, the imprecise parameter estimates of β_0 and β_2 partially offset each other as seen in the bottom of Figure IV.9 where the estimated kernel (red) is only slightly below the true kernel (black) when accounting for the baseline effect. This indicates some correlation effect between β_0 and β_2 when estimating the levels.

Figure IV.10 displays the decoding of the true states (black) by the mean of the 50 particles (red), for the last 250 ms of the data. The true burst periods are captured relatively precise, but some errors especially around exiting bursts.

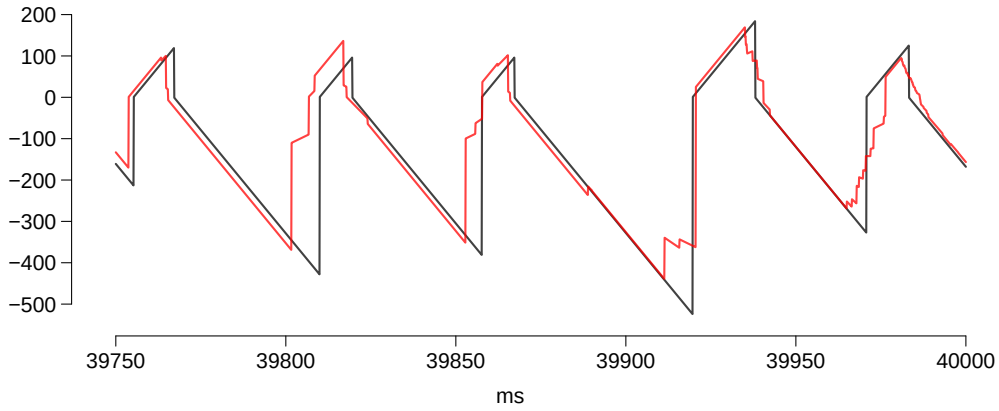

FIGURE IV.10: The last 250 ms of the decode X_k process (red) versus the true simulated process (black). The overall decoding of bursting/resting is relatively precise with 50 particles.

Figure IV.11 shows the last 250 ms of the estimated intensity (orange) with the observed spikes (black) and the probability of the neuron bursting (red). The latter was defined as the proportion of the 50 particles with positive state at time t_k. As in Figure IV.10, it is evident that the particle filter decodes the bursts quite well, but with some uncertainty, not surprisingly, when entering/exiting bursts. Nevertheless the intensity exhibits rapid increase/decrease at the time of spikes. Especially around the endpoints of bursts, this is much more pronounced than what was observed in Østergaard, Kramer, and Eden (2018) which use a GLM without a latent state process controlling bursting behavior. Thus, including the process X_k in the model, has a very positive effect concerning the estimation of the intensity process. It should be noted, that the model specification used here, where the neuron was assumed not to spike when resting by setting $h^r = \beta_0^r = 0$, ensured a simpler identification of the latent state. With this assumption, when a spike is observed, the neuron must be in bursting mode and whenever there is a longer delay between spikes, the filter interprets this as a resting period. This implies that one can view the latent process X_k

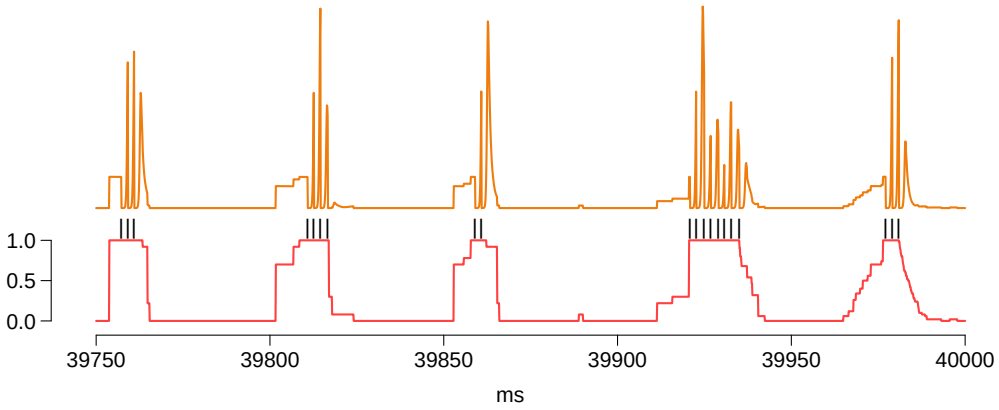

FIGURE IV.11: Last 250 ms of the estimated intensity (orange) compared to observed spikes (black) and the estimated probability of bursting (red), based on $M = 50$ particles.

as partially observed, which is easily seen when plotting the spike train as in Figure IV.2. Here is it relatively obvious when the neuron is bursting and when it is not. However, more complex cases, where it would be difficult to assess by eyeballing, exist of course, also making it more difficult for the particle filter to decode the bursting/resting periods. One such case could be to determine whether two spikes belong to the same burst or not, when the period between is longer than, say, 20 ms at the right tail of the burst length distribution. In this case, with the transition probabilities in Figure IV.4, the decision on bursting/resting becomes ambiguous.

To determine how well the overall model fit the data, we also performed a Kolmogorov-Smirnov (KS) test on the rescaled spike times versus an $\mathrm{Exp}(1)$ distribution (see Kass, Eden, and Brown, 2014). With a KS statistic of 0.036 and a corresponding p-value of 0.085, these values implied that the estimated model explains the data well. Figure IV.12 displays the corresponding Kolmogorov-Smirnov plot, which shows that empirical distribution of rescaled spike times versus the theoretical $\mathrm{Exp}(1)$ (in red) agrees and the discrepancy is within approximate 95% error bounds (blue).

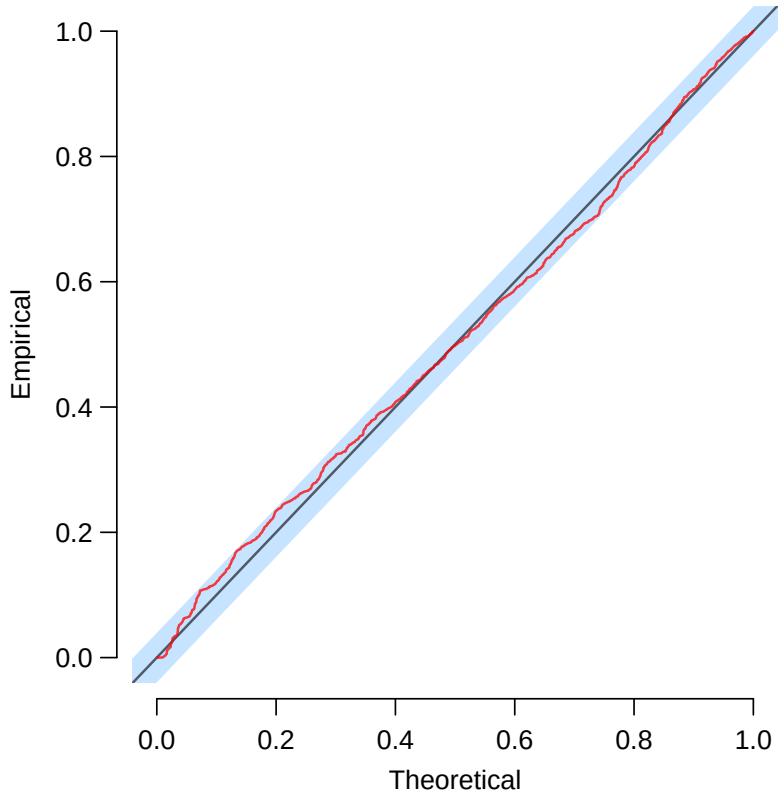

FIGURE IV.12: Kolmogorov-Smirnov plot for the SSGLM neuron. The Kolmogorov-Smirnov statistic is 0.036 with a p-value 0.085, indicating a good fit of the model. This is also evident from the comparison of the empirical and theoretical distributions (red), which is within the approximate 95% boundaries (blue).

5.3 Analysis of a bursting Izhikevich neuron

We performed an analysis for a bursting Izhikevich neuron, analogously to the one above for the SSGLM neuron, with the similar spline knots and transition probabilites. The parameter convergence is shown in Figure IV.13, where it is clear that the

baseline parameter β_0 converges to around -1 and β_2 converges to a level between 1 and 1.5. As for the SSGLM neuron, the β_3 parameter exhibits a less steady convergence, but fluctuates somewhat around -0.5. The data was also recycled to obtain a total of 40 s of observations, but it does not appear to be necessary as parameters β_0 and β_2 converge already at 20 s, and β_3 does not show a much more stable convergence after 20 s. Figure IV.14 presents the estimated kernel without the base-

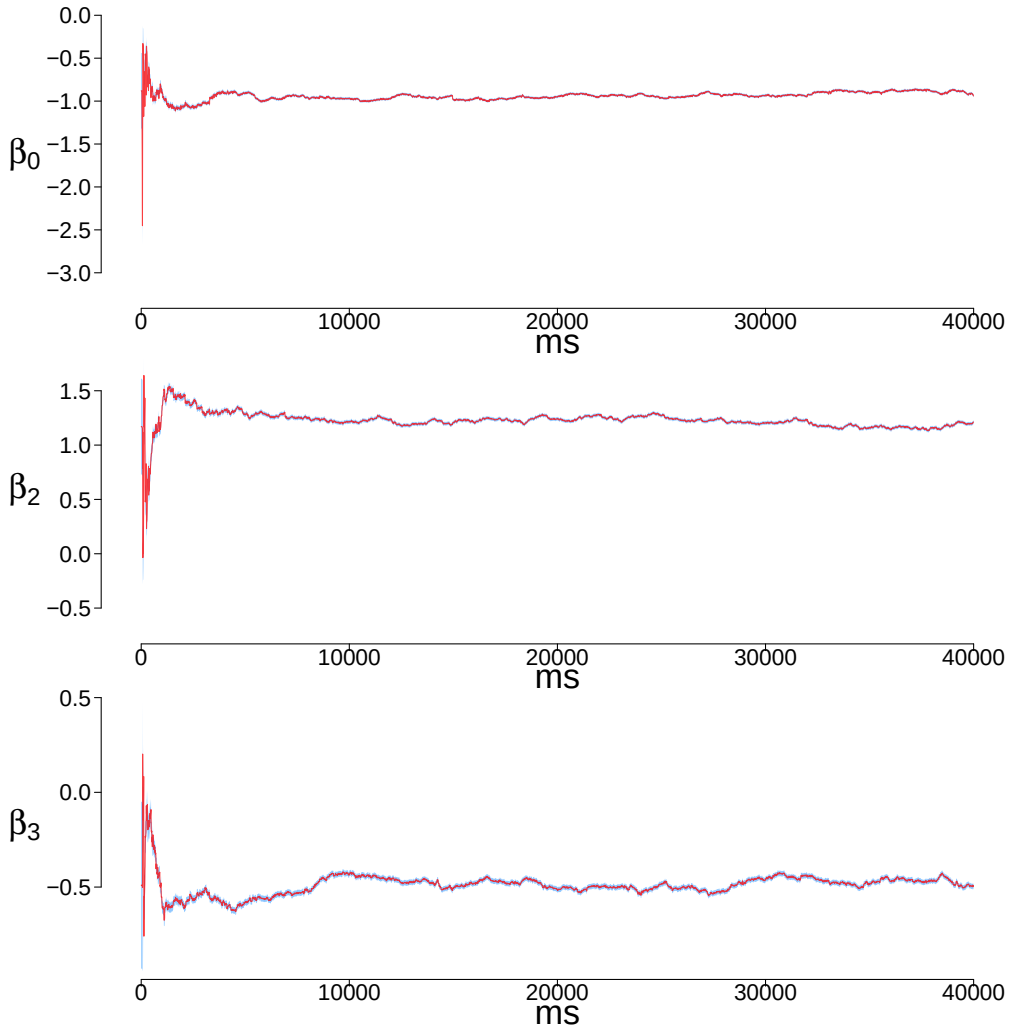

FIGURE IV.13: Convergence of parameters for the Izhikevich neuron. Parameter estimates in red. Note that the data has been repeated 2x.

line effect (in the exponential domain), based on parameter estimates $\hat{\beta}_0 = -0.90$, $\hat{\beta}_2 = 1.18$ and $\hat{\beta}_3 = -0.47$. The simulation kernel for the SSGLM neuron was tuned to mimic a bursting Izhikevich neuron, hence the very similar shapes. The absolute refractory period of the (Izhikevich) neuron at 0-1.5 ms is clearly visible in Figure IV.14, due to the fixed parameter $\beta_1 = -10$. The probability of spiking then increases when a spike is observed up to ≈ 4 ms in the past, and for spikes between 4-10 ms in the past implies a slightly suppressed spiking probability. Since the kernel ends at 0 (1 in the exponential domain), this implies that there is not more influence from past spikes at 10 ms and beyond.

Figure IV.15 shows the estimated intensity (orange) with observed spikes (black) and the estimated probability of whether the neuron is bursting or not (red). As

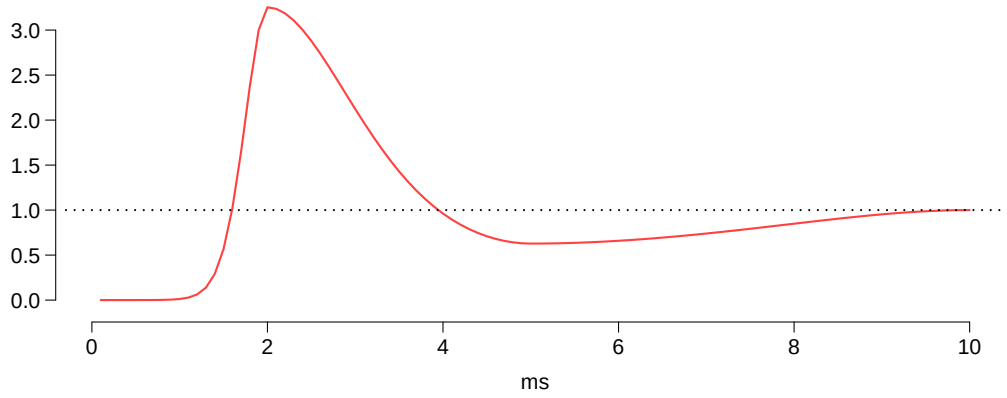

FIGURE IV.14: Estimated bursting kernel for the Izhikevich neuron in the exponential domain, based on the last 10 s.

in the case of the SSGLM neuron, the intensity increase/decrease fast when entering/exiting a burst. The decoding of the latent state shows a tendency to overesti-

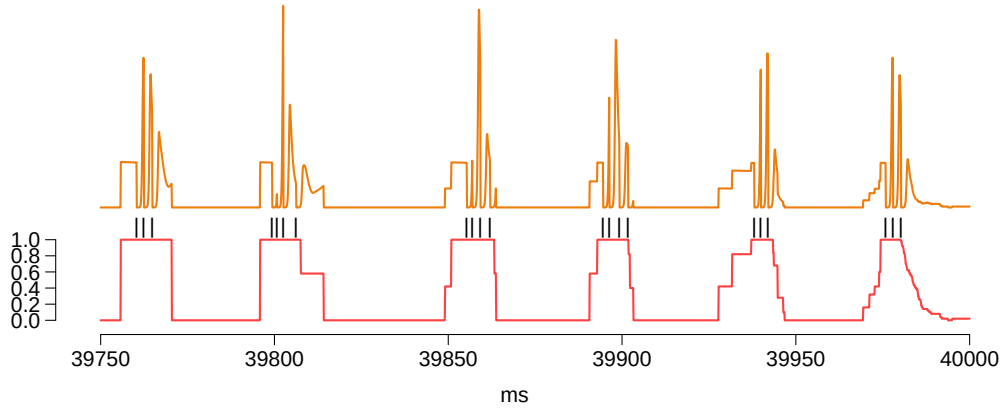

FIGURE IV.15: Last 250 ms of the estimated intensity (orange) for the Izhikevich neuron, compared to observed spikes (black) and the estimated probability of bursting (red), based on $M = 50$ particles.

mate the length bursts, but taking the estimated intensity into account, the peaks of the intensity agrees with the observed spike times. Hence, considering the combined decoding and parameter estimation of the MPFPP algorithm, the overall estimation is very good. This is also evident from the Kolmogorov-Smirnov statistic of 0.031, with a corresponding p-value of 0.078. The complementary KS plot is presented in Figure IV.16. Both the statistic and Figure IV.16 indicates that the model fits the data very well. This is in contrast to the findings in Østergaard, Kramer, and Eden (2018), where the KS statistic was found to be significant when comparing the rescaled spike times to the $\mathrm{Exp}(1)$ distribution. In that paper, the model formulation used only the GLM part, including a kernel, whereas here, we have also included the latent process X_k. Figure IV.16 supports the need for such a latent state in the model formulation, in order to better capture the dual behavior of bursting (Izhikevich) neurons.

5.4 Probability of bursting versus Izhikevich variables

As a final interesting comparison, we investigated the connection between the u_t, v_t variables of the Izhikevich model (IV.20) and the decoding property of the SSGLM.

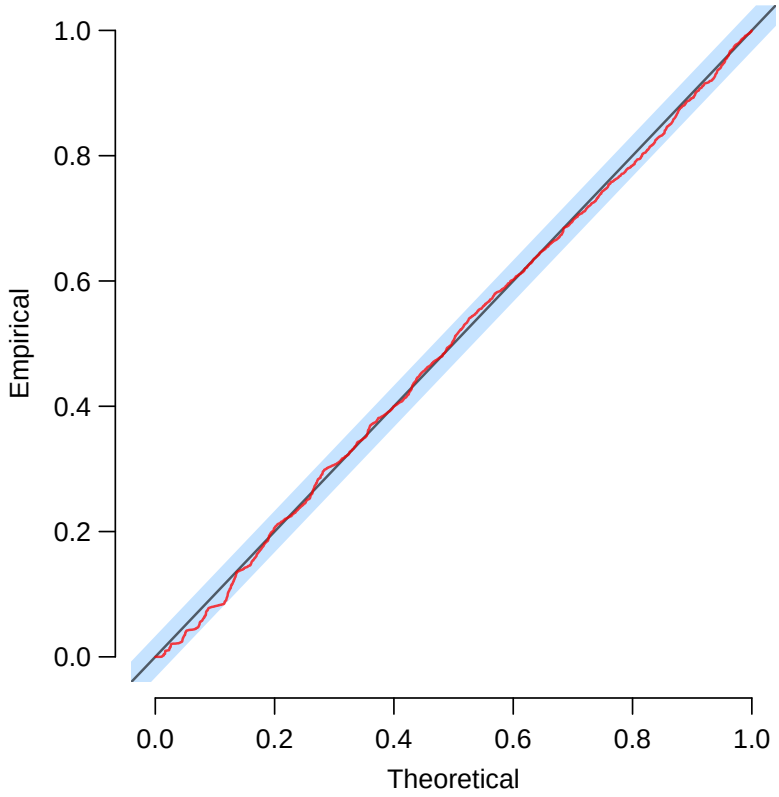

FIGURE IV.16: KS plot for the Izhikevich neuron. The KS statistic is 0.031 with a p-value 0.078, indicating a good fit of the model.

Figure IV.17 displays the simulated Izhikevich model in the u_t, v_t-plane as dots, but colored with the decoded probability of bursting based on the estimated SSGLM, using the MPFPP algorithm. It is clear that when $v_t < -50$mV (the resetting level c) the probability of bursting is close to 0 (blue) whereas for $v_t > -50$ the probability approaches 1 (red). As the neuron enters bursting, the probability of bursting changes quickly to 1, and stays at 1 until the u_t, v_t variables cross the separatrix (black) of the deterministic Izhikevich neuron (orange). This indicate a link between the Izhikevich variables, and the latent state of the SSGLM and could be of interest to study this link in detail also for other higher dimensional dynamical models with more than two variables, capable of bursting. In relation to the Izhikevich model, the resetting parameter d, which bumps the u_t variable when a spike is observed, controls the number of spikes in a burst. As such, there must be an intrinsic association between the transition probabilities of the SSGLM and the d parameter. This parameter is also of interest in relation to the level of noise (σ) in the simulation of the Izhikevich model. Since changing the noise level impacts the u_t variable indirectly, it is apparent from Figure IV.17 that the number of spikes becomes more varied as the uncertainty in crossing the separatrix increases. In the deterministic model, the number of spikes is always five, but adding a bit of noise will only influence whether the fifth spike is included or not, hence a bimodal burst length distribution becomes visible. If σ is then increased, the noise begins to affect whether both the fourth and the fifth spike is included in a burst. Hence, the noise level controls the variability in the burst lengths, in terms of the number of spikes present.

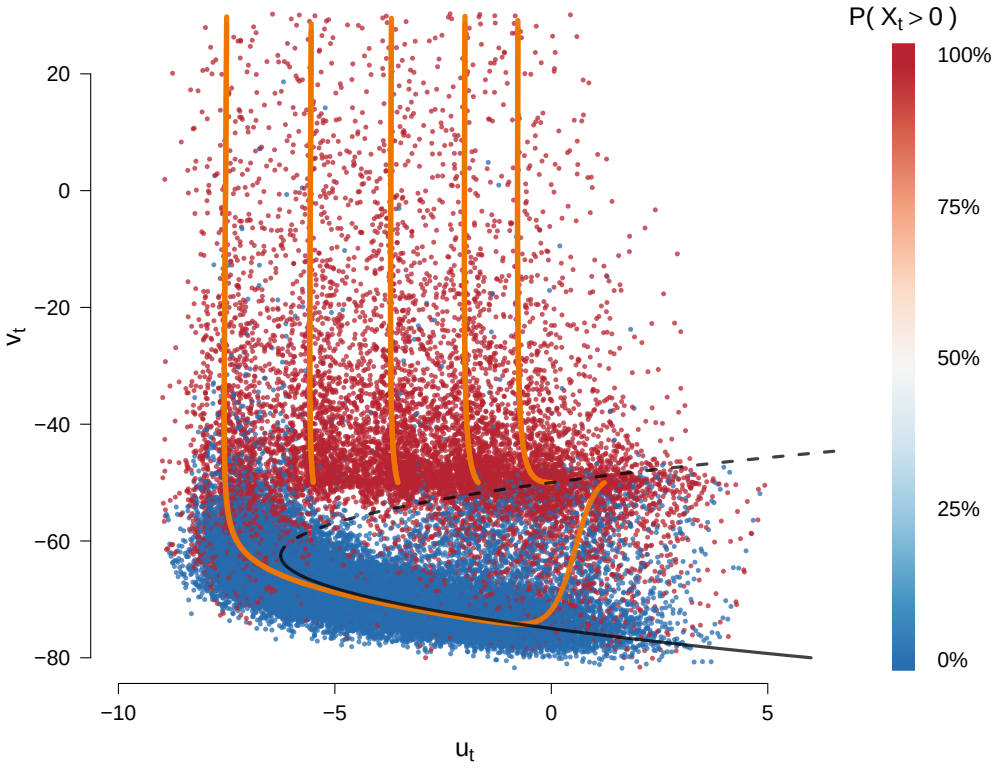

FIGURE IV.17: The (u_t, v_t)-plane of the simulated bursting Izhikevich neuron. Observations are represented as colored dots, where the color corresponds to the MPFPP decoded latent state of bursting/resting. The separatrix show the stable (black) and unstable (black dashed) boundaries for the deterministic Izhikevich model (orange).

6 Discussion

This paper addressed the fact that a bursting neuron exhibit two modes: bursting and resting. As highlighted in Østergaard, Kramer, and Eden (2018) this property was not adequately captured by a multiplicatively separable history dependent GLM. Here, we have proposed a possible remedy for this issue, by including a separate mechanism that controls for bursting behavior. Specifically we extended the GLM from Østergaard, Kramer, and Eden (2018) with a latent state that switches between bursting and resting, thus explicitly accounting for this property. In addition to this, the history dependency was modeled using spline basis functions, to reduce the number of parameters, ultimately leading to a simple, yet flexible model formulation using only a few parameters. In essence, our model formulation includes three main components:

- Two baseline firing rates.
- Two kernels that modulates the baseline firing rates, based on previous observed spikes.
- A latent state controlling which of the baselines/kernels are effective, at a given timepoint.

By assuming the resting baseline and kernel parameters to be zero, the estimation only concerned parameters for the bursting mode, as the transition probabilities that models the switching behavior were kept fixed. The latter were estimated in a preliminary study on a bursting Izhikevich neuron. However, including these in the estimation is possible and would make an interesting study on various noise levels for the Izhikevich neuron, as well as other generating models and real data. The simple structure of the model ensures a straightforward interpretation of the components, evident from the two analyses in this paper.

The estimation was performed with a sequential Monte Carlo approach by adapting a marginalized particle filter algorithm for use on point process observations. The algorithm was shown to be very effective in partitioning the state space into a linear and nonlinear part, requiring fewer particles at each timestep. As mentioned above, the estimation did not include the parameters of the transition probabilities, but including these in the marginalized particle filter would make this model suitable for a bursting Izhikevich neuron with varying noise input. In this case, the model presented in this paper would also be applicable for real time observations, as the estimation of transition probabilities would imply a necessary adaptive property. However, for a given data set of bursting neurons, performing an exploratory analysis would enable fixation of these parameters, but never the less, this leaves the overall solution less generic. As such, analyzing an estimation where the transition probabilities can adapt over time would make an interesting project. A possible extension to the model presented here would be to account for the adaptiveness of spiking observed in neurons. If the latent state is known (decoded), one can formulate a modified kernel, which depends directly on value of the latent state, not only the sign.

Previous work has shown that the GLM class is capable of modeling a wide variety of neuron types (Weber and Pillow, 2017; Østergaard, Kramer, and Eden, 2018) as is the case for the Izhikevich model. The model presented here included relatively few parameters, similar to the Izhikevich model. However, contrary to the (deterministic) Izhikevich model, the model proposed here is a stochastic model which can quantify the variability in (real) observed data and proper methods exist to assess how an estimated model fits the observations, such as the Kolmogorov-Smirnov statistic and plot, used here. As the model was shown capable of replicating the behavior of an Izhikevich model as well as fitting well to Izhikevich observations, the capabilities of the SSGLM, compared to the Izhikevich model, are suitable for experimental use such as in classifying neurons based on their intrinsic properties. Various types of bursting could be identified not only through the bursting kernel, but also the transition probabilities. Non-bursting neurons, such as tonic spiking, could be identified by having "bursts" with only a single spike.

In conclusion, the SSGLM can be seen as a flexible model, capable of capturing complex spike train pattern such as bursting, while at the same time keeping the number of parameters at a minimum and the estimated model is easily interpreted. This could be of interest for experimental researchers in need of a model that can decode dual behavior in observed neurons, and possibly classify neurons according to some pre-specified criteria, such as bursting.

A Estimation of transition probabilites

To estimate the transition probabilities from the observed Izhikevich neuron, define Z as either the length of a burst or a resting period. Here we let Z be the length of a burst, but the following arguments can be applied analogously to the resting period. Defining Z as the length of a burst implies that

$$\begin{aligned}
P(Z = z) &= P(X_{k+1} = 0|\, X_k = z, X_{k-1} = z-1, \ldots, X_{k-z+1} = 1) \\
&= P(X_{k+1} = 0|\, X_k = z) P(X_k = z|\, X_{k-1} = z-1) \cdots P(X_{k-z+1} = 1) \\
&= P(X_{k+1} = 0|\, X_k = z) \prod_{i=2}^{z} P(X_{k-z+i} = i|\, X_{k-z+i-1} = i-1) P(X_{k-z+1} = 1) \\
&= \big(1 - g^b(z; \alpha, \kappa)\big) \prod_{i=2}^{z} g^b(i-1; \alpha, \kappa) \\
&= \big(1 - g^b(z; \alpha, \kappa)\big) \prod_{i=1}^{z-1} g^b(i; \alpha, \kappa) \\
&= \frac{\alpha^\kappa}{\alpha^\kappa + z^\kappa} \prod_{i=1}^{z-1} \frac{i^\kappa}{\alpha^\kappa + i^\kappa}
\end{aligned} \tag{IV.25}$$

where we have defined $P(X_{k-m+1} = 1) = 1$ as it is assumed that the neuron is bursting and

$$g^b(z; \alpha, \kappa) = \frac{z^\kappa}{\alpha^\kappa + z^\kappa} = P(X_{k+1} = z+1 | X_k = z)$$

are the transition probabilities modeled as log-logistic functions. Parameters α and κ were then fitted using (IV.25) by least squares to the empirical log-proportions of observed burst lengths using optimization in R, R Core Team (2017).

Conclusion and Outlook

CHAPTER 4

Conclusion

1 Discussion

With this thesis we have investigated two different areas concerning the modeling of neural data. Part one examined a new way of inferring the network structure of coupled neurons. Although the applied methodology is not new it appears that Manuscript I is the first succesful application of cointegration analysis in a neuroscience context. Here a network structure was analyzed through the phase processes of the neurons and it was demonstrated that cointegration analysis is a promising area for future research possibilities for coupled neurons. Continuing down the cointegration path, Manuscript II addressed the question of how cointegration analysis may be applied for high dimensional data. This work is not yet finalized, but it shows supportive evidence of using cointegration for this type of data as well. Presently a major outstanding task is to develop a solid theoretical framework, where the classical linear restrictions for the parameters of cointegration models may be extended to include information on sparseness structure, as well as other structured couplings. In the working paper we have only demonstrated the potential for a linear coupling, inspired by the Kuramoto model, with a symmetric structure. However, it was shown that even such a simple specification is not trivial to impose in a cointegration estimation procedure when considering both symmetry and reduced rank constraints. This also calls for new results in cointegration, such as theoretically based numerical methods for penalization in cointegration models. A current drawback for cointegration analysis is the lack of an intuitive understanding of how cointegration influences the neurons, besides showing that the structure changes when the couplings in a network change. If this is to become a popular technique in the neuroscience community, a clearcut interpretation of cointegration parameters in connection to the network coupling is necessary. In order to obtain this goal it would require analyzing more experimental data of real neurons. However, since more and more data is becoming available this should be a realizable task.

Part two of this thesis examined the applicability of Generalized Linear Models (GLM) for spike train data analysis. Manuscript III investigated the flexibility of a simple GLM, when used to capture various types of spike patterns also for varying noise levels of the data generating process. The paper showed that the GLM is a suitable choice for different spike patterns, but also that a certain amount of noise is preferable when fitting a model of a stochastic nature, such as the GLM. One of

the shortcomings of the GLM that were highlighted in this paper was addressed in Manuscript IV, where the GLM formulation of the previous paper was extended to a state space model, thus accounting for the specific behavior of bursting neurons. This demonstrated that extending the GLM for specific purposes further increases the flexibility of this model class, and that it is possible to construct a statistical model that is very capable of similar behavior as the well accepted Izhikevich model. These two papers can be seen as making a case for statistical modeling of neural data. The usefulness of specific statistical models have been demonstrated with emphasis on estimation procedures based on well known results. As in the case of cointegration analysis, a shortcoming for statistical models is the explanation of model parameters. By developing a stronger link between well understood and researched models, such as the Izhikevich, but also other classical models of neurons, statistical models such as the GLM and it's extensions could be further accepted as the de facto standard concerning analysis of neural data.

2 Extensions

There are a number of possibilities to extend the work presented here. As mentioned in the discussion above an interesting extension to the work on cointegration would be to expand the high dimensional framework to include sparsity constraints, as well as structured restrictions other than symmetry. This could be useful for analyzing data of coupled clusters where the clusters are coupled not only internally, but couplings also exist externally between clusters. Another new direction would be to work on models with non-linear phase coupling. This could include a switch that turns cointegration on/off or reverse the effect. The latter is necessary in order to build a cointegration model that approximates the actual Kuramoto model, and not just a linear version of it. It is evident from the fact that linearizing the sine function amounts to a sign of ± 1, depending on where the function is linearized. It would be an interesting problem to overcome, but would require some theoretical work to expand the framework presented here. Finally, a compelling idea would be to investigate any possible connections between point process models, such as the ones analyzed in the second part of this thesis, and cointegration. A vague direction of this objective could be to look into multivariate point processes where the intensity was allowed to cointegrate. However, it is not currently clear how, or whether this is at all sensible, to write such a model, but being able to apply cointegration techniques for point processes would be a very attractive extension.

3 Conclusion

The aim of the research projects presented here concerned the analysis of neural data from different perspectives. As we have shown, using analysis techniques common to the realm of econometrics can help to gain insight into the structures of neural networks, and modeling event times with various patterns could help to classify types of observed neurons. While statistical modeling is gaining acceptance in the community of neuroscience researchers, the models considered here are not only useful in this context. Investigating the use of cointegration in a high dimensional context is useful in relation to big data, where the dimension of data increases and the quest

to find relational structure becomes increasingly difficult. By researching new statistical tools for this type of analysis, which are applicable to high dimensional data, we can increase our understanding of the mechanisms in the brain, such as the connectivity of it's internal components. Although we might never completely unveil the enigmas of the brain we can certainly attempt to construct a continuously better explanation of it's capabilities. Hopefully this thesis has assisted in this quest by suggesting new statistical models that may be of use for experimental researchers to answer some of their questions.

APPENDIX A

Programming

1 Overview

This appendix contain information on the programs used in the thesis. The code is distributed in three `R` packages, R Core Team, 2017.

Package name	Usage
`cods`	Used in the manuscript on cointegrated oscillators
`BU`	Used in the manuscript on GLMs for Izhikevich neurons
`particleGLM`	Used for parameter estimation of the bursting neuron model

These packages are currently[1] not publicly available, but they are planned for future publication. As such, they are available upon request.

2 R package: `cods`

This package was created to meet the requirements for a continuous time cointegration analysis. As such it performs standard cointegration estimation, but with the inclusion of a time observation frequency. Core elements are implemented in C++ with wrapper functions in R scripts.

Main functions include `johansen` and `bootstrap` which performs estimation and rank test respectively, both are written as wrappers to the C++ implementations. Auxiliary functions include calculation of standard errors (`getSE`), Hilbert transform (`hilbert`), likelihood ratio test (`LRtest`) and the mean phase coherence measure (`meanPhaseCoherence`).

3 R package: `BU`

This package was created as a collection of scripts and C++ functions for the study on Generalized Linear Models and Izhikevich neurons. Main elements include simulation of Izhikevich neurons (`izhikevich`) and fitting GLMs (`doFit`). Auxiliary elements include setting up spline basis functions (`make_splines`) and applying the

[1] At the time of writing, December 2017.

filter to the spike trains (`filter`) as well as calculating the input to the Kolmogorov-Smirnov plot (`doKS`).

4 R package: `particleGLM`

This package implements the marginalized particle filter for point processes (MPFPP). The MPFPP is implemented in C++ (`mpfpp`). Some auxiliary elements include calculating burst lengths (`burst_length`), the transition probabilites (`px_x`) and the emission probabilities (`py_xy`).

Bibliography

Ahmadian, Yashar, Jonathan W. Pillow, and Liam Paninski (2011). "Efficient Markov Chain Monte Carlo Methods for Decoding Neural Spike Trains". In: *Neural Comput.* 23.1, pp. 46–96.

Ali, Mohsin M., Kristin K. Sellers, and Flavio Fröhlich (2013). "Transcranial Alternating Current Stimulation Modulates Large-Scale Cortical Network Activity by Network Resonance". In: *Journal of Neuroscience* 33.27, pp. 11262–11275.

Ashwin, Peter, Stephen Coombes, and Rachel Nicks (2016). "Mathematical Frameworks for Oscillatory Network Dynamics in Neuroscience". In: *The Journal of Mathematical Neuroscience* 6.1, pp. 1–92.

Bec, Frédérique and Anders Rahbek (2004). "Vector equilibrium correction models with non-linear discontinuous adjustments". In: *The Econometrics Journal* 7.2, pp. 628–651.

Boswijk, H. Peter et al. (2016). "Inference on co-integration parameters in heteroskedastic vector autoregressions". In: *Journal of Econometrics* 192.1, pp. 64–85.

Brown, Emery N. et al. (2002). "The Time-rescaling Theorem and Its Application to Neural Spike Train Data Analysis". In: *Neural Comput.* 14.2, pp. 325–346.

Brunel, Nicolas and Peter E. Latham (2003). "Firing Rate of the Noisy Quadratic Integrate-and-Fire Neuron". In: *Neural Computation* 15.10, pp. 2281–2306.

Brunel, Nicolas and Mark C. W. van Rossum (2007). "Lapicque's 1907 paper: from frogs to integrate-and-fire". In: *Biological Cybernetics* 97.5, pp. 337–339.

Burton, Shawn D., G. Bard Ermentrout, and Nathaniel N. Urban (2012). "Intrinsic heterogeneity in oscillatory dynamics limits correlation-induced neural synchronization". In: *Journal of Neurophysiology* 108.8, pp. 2115–2133.

Cappé, Olivier, Eric Moulines, and Tobias Ryden (2005). *Inference in Hidden Markov Models (Springer Series in Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc. ISBN: 0387402640.

Cavaliere, Giuseppe, Anders Rahbek, and A. M. Robert Taylor (2012). "Bootstrap Determination of the Co-Integration Rank in Vector Autoregressive Models". In: *Econometrica* 80.4, pp. 1721–1740.

Chu, Moody T., Robert E. Funderlic, and Robert J. Plemmons (2003). "Structured low rank approximation". In: *Linear Algebra and its Applications* 366.Supplement C. Special issue on Structured Matrices: Analysis, Algorithms and Applications, pp. 157 –172.

Dahlhaus, R and J Neddermeyer (2012). "On The Relationship Between The Theory Of Cointegration And The Theory Of Phase Synchronization". Working paper: arXiv:1201.0651.

Dimatteo, Ilaria, Christopher R. Genovese, and Robert E. Kass (2001). "Bayesian curve fitting with free knot splines". In: *Biometrika* 88.4, p. 1055.

Doucet, Arnaud and Adam M Johansen (2009). "A tutorial on particle filtering and smoothing: Fifteen years later". In: *Handbook of nonlinear filtering* 12.656-704, p. 3.

Eckart, Carl and Gale Young (1936). "The approximation of one matrix by another of lower rank". In: *Psychometrika* 1.3, pp. 211–218.

Eden, Uri T. et al. (2004). "Dynamic Analysis of Neural Encoding by Point Process Adaptive Filtering". In: *Neural Computation* 16.5, pp. 971–998.

Engle, Robert F. and C. W. J. Granger (1987). "Co-Integration and Error Correction: Representation, Estimation, and Testing". In: *Econometrica* 55.2, pp. 251–276.

Ermentrout, G. Bard (1985). "Synchronization in a pool of mutually coupled oscillators with random frequencies". In: *Journal of Mathematical Biology* 22.1, pp. 1–9.

Fan, Ky and A. J. Hoffman (1955). "Some Metric Inequalities in the Space of Matrices". In: *Proceedings of the American Mathematical Society* 6.1, pp. 111–116.

Fernandez, Bastien and Lev S. Tsimring (2014). "Typical trajectories of coupled degrade-and-fire oscillators: from dispersed populations to massive clustering". In: *Journal of Mathematical Biology* 68.7, pp. 1627–1652.

FitzHugh, Richard (1961). "Impulses and Physiological States in Theoretical Models of Nerve Membrane". In: *Biophysical Journal* 1.6, pp. 445–466.

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software* 33.1, pp. 1–22.

Geer, Sara van de et al. (2014). "On asymptotically optimal confidence regions and tests for high-dimensional models". In: *Ann. Statist.* 42.3, pp. 1166–1202.

Granger, C.W.J. (1981). "Some properties of time series data and their use in econometric model specification". In: *Journal of Econometrics* 16.1, pp. 121–130.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

Herculano-Houzel, Suzana (2009). "The Human Brain in Numbers: A Linearly Scaled-up Primate Brain". In: *Frontiers in Human Neuroscience* 3, p. 31.

Hertäg, Loreen, Daniel Durstewitz, and Nicolas Brunel (2014). "Analytical approximations of the firing rate of an adaptive exponential integrate-and-fire neuron in the presence of synaptic noise". In: *Frontiers in Computational Neuroscience* 8, p. 116.

Hodgkin, A L and A F Huxley (1952). "A quantitative description of membrane current and its application to conduction and excitation in nerve". In: *The Journal of Physiology* 117.4, pp. 500–544.

Izhikevich, Eugene M. (2003). "Simple Model of Spiking Neurons". In: *Trans. Neur. Netw.* 14.6, pp. 1569–1572.

— (2004). "Which Model to Use for Cortical Spiking Neurons?" In: *Trans. Neur. Netw.* 15.5, pp. 1063–1070.

— (2010). *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting*. Computational Neuroscience. MIT Press.

Izhikevich, Eugene M. and Gerald M. Edelman (2008). "Large-scale model of mammalian thalamocortical systems". In: *Proceedings of the National Academy of Sciences* 105.9, pp. 3593–3598.

Johansen, Søren (1996). *Likelihood-Based Inference in Cointegrated Vector autoregressive Models*. Oxford University Press.

Johansen, Søren and Ernst Schaumburg (1999). "Likelihood Analysis of Seasonal Cointegration". In: *Journal of Econometrics* 88.2. JEL classification codes: C32, pp. 301–339.

Kalman, R. E. (1960). "A New Approach to Linear Filtering and Prediction Problems". In: *Journal of Basic Engineering* 82.1, pp. 35–45.

Kass, R. E. and V. Ventura (2001). "A Spike-Train Probability Model". In: *Neural Computation* 13.8, pp. 1713–1720.

Kass, Robert E., Uri T. Eden, and Emery N. Brown (2014). *Analysis of Neural Data*. Springer.

Kessler, Mathieu and Anders Rahbek (2001). "Asymptotic Likelihood Based Inference for Co-integrated Homogenous Gaussian Diffusions". In: *Scandinavian Journal of Statistics* 28.3, pp. 455–470.

Kessler, Matthieu and Anders Rahbek (2004). "Identification and Inference for Multivariate Cointegrated and Ergodic Gaussian Diffusions". In: *Statistical Inference for Stochastic Processes* 7.2, pp. 137–151.

Kristensen, Dennis and Anders Rahbek (2013). "Testing and Inference in Nonlinear Cointegrating Vector Error Correction Models". In: *Econometric Theory* 29.6, pp. 1238–1288.

Kuramoto, Y (1984). *Chemical Oscillations, Waves and Turbulence*. Springer.

Lapicque, Louis (1907). "Recherches quantitatives sur l'excitation électrique des nerfs traitée comme une polarisation". In: *Journal De Physiologie Et De Pathologie Generale* 9, pp. 620–635.

Latimer, Kenneth W. et al. (2014). "Inferring Synaptic Conductances from Spike Trains Under a Biophysically Inspired Point Process Model". In: *Proceedings of the 27th International Conference on Neural Information Processing Systems*. NIPS'14. Montreal, Canada: MIT Press, pp. 954–962.

Lütkepohl, Helmut (2005). *New Introduction to Multivariate Timeseries Analysis*. Springer.

Ly, Cheng (2014). "Dynamics of Coupled Noisy Neural Oscillators with Heterogeneous Phase Resetting Curves". In: *SIAM Journal on Applied Dynamical Systems* 13.4, pp. 1733–1755.

Ly, Cheng and G. Bard Ermentrout (2011). "Analytic approximations of statistical quantities and response of noisy oscillators". In: *Physica D Nonlinear Phenomena* 240.8, pp. 719–731.

Macke, Jakob H et al. (2011). "Empirical models of spiking in neural populations". In: *Advances in Neural Information Processing Systems 24*. Ed. by J. Shawe-Taylor et al. Curran Associates, Inc., pp. 1350–1358.

Martens, Erik (2013). "Chimera states in mechanical oscillator networks". In: *Proceedings of the National Academy of Sciences* 110.26, pp. 10563–10567.

McCullagh, Peter and John A Nelder (1989). *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC.

Mormann, Florian et al. (2000). "Mean phase coherence as a measure for phase synchronization and its application to the {EEG} of epilepsy patients". In: *Physica D: Nonlinear Phenomena* 144.3-4, pp. 358–369.

Morris, C and H Lecar (1981). "Voltage oscillations in the barnacle giant muscle fiber." In: *Biophysical Journal* 35.1, pp. 193–213.

Nadim, Farzan et al. (1995). "Modeling the leech heartbeat elemental oscillator I. Interactions of intrinsic and synaptic currents". In: *Journal of Computational Neuroscience* 2.3, pp. 215–235.

Nagumo, J., S. Arimoto, and S. Yoshizawa (1962). "An Active Pulse Transmission Line Simulating Nerve Axon". In: *Proceedings of the IRE* 50.10, pp. 2061–2070.

Nelder, J. A. and R. W. M. Wedderburn (1972). "Generalized Linear Models". In: *Journal of the Royal Statistical Society. Series A (General)* 135.3, pp. 370–384.

Øksendahl, Bernt (2007). *Stochastic Differential Equations*. Springer.

Oliveira, Henrique M. and Luís V. Melo (2015). "Huygens synchronization of two clocks". In: *Scientific Reports* 5, 11548 EP –.

Østergaard, Jacob, Mark A. Kramer, and Uri T. Eden (2018). "Capturing Spike Variability in Noisy Izhikevich Neurons Using Point Process Generalized Linear Models". In: *Neural Computation* 30.1. PMID: 29064782, pp. 125–148.

Østergaard, Jacob, Anders Rahbek, and Susanne Ditlevsen (2017a). "Cointegration analysis of High-Dimensional linear Kuramoto networks". Working paper.

— (2017b). "Oscillating systems with cointegrated phase processes". In: *Journal of Mathematical Biology* 75.4, pp. 845–883.

Østergaard, Jacob et al. (2017). "State Space Model for Bursting Neurons". Working paper.

Paninski, Liam (2006). "The Spike-Triggered Average of the Integrate-and-Fire Cell Driven by Gaussian White Noise". In: *Neural Computation* 18.11, pp. 2592–2616.

Pikovsy, Arkady, Michael Rosenblum, and Jürgen Kurths (2001). *Synchronization: A Universal Concept in Nonlinear Sciences*. Cambridge University Press.

Pillow, Jonathan W. et al. (2008). "Spatio-temporal correlations and visual signalling in a complete neuronal population". In: *Nature* 454.7207, pp. 995–999.

Prinz, Astrid A., Cyrus P. Billimoria, and Eve Marder (2003). "Alternative to Hand-Tuning Conductance-Based Models: Construction and Analysis of Databases of Model Neurons". In: *Journal of Neurophysiology* 90.6, pp. 3998–4015.

Prinz, Astrid A, Dirk Bucher, and Eve Marder (2004). "Similar network activity from disparate circuit parameters". In: *Nat Neurosci* 7.12, pp. 1345–1352.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Sarma, Sridevi V et al. (2012). "The effects of cues on neurons in the basal ganglia in Parkinson's disease". In: *Frontiers in Integrative Neuroscience* 6, p. 40.

Schmidt, Torben, Søren Johansen, and Peter Thejll (2012). "Statistical Analysis of Global Surface Temperature and Sea Level Using Cointegration Methods". In: *Journal of Climate* 25.22, pp. 7822–7833.

Schon, T., F. Gustafsson, and P. J. Nordlund (2005). "Marginalized particle filters for mixed linear/nonlinear state-space models". In: *IEEE Transactions on Signal Processing* 53.7, pp. 2279–2289.

Shoeb, Ali H (2009). "Application of Machine Learning to Epileptic Seizure Onset Detection and Treatment". Dissertation. MIT Health Sciences and Technology Division.

Strogatz, Steven H. (1987). "Human sleep and circadian rhythms: a simple model based on two coupled oscillators". In: *Journal of Mathematical Biology* 25.3, pp. 327–347.

— (2000). "From Kuramoto to Crawford exploring the onset of synchronization in populations of coupled oscillators". In: *Physica D Nonlinear Phenomena* 143.1-4, pp. 1–20.

Taylor, David and Philip Holmes (1998). "Simple models for excitable and oscillatory neural networks". In: *Journal of Mathematical Biology* 37.5, pp. 419–446.

Taylor, Jonathan and Robert J. Tibshirani (2015). "Statistical learning and selective inference". In: *Proceedings of the National Academy of Sciences* 112.25, pp. 7629–7634.

Tibshirani, Robert (1996). "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1, pp. 267–288.

Traub, R. D. et al. (1991). "A model of a CA3 hippocampal pyramidal neuron incorporating voltage-clamp data on intrinsic conductances". In: *Journal of Neurophysiology* 66.2, pp. 635–650.

Truccolo, Wilson et al. (2005). "A Point Process Framework for Relating Neural Spiking Activity to Spiking History, Neural Ensemble, and Extrinsic Covariate Effects". In: *Journal of Neurophysiology* 93.2, pp. 1074–1089.

Vanier, Michael C. and James M. Bower (1999). "A Comparative Survey of Automated Parameter-Search Methods for Compartmental Neural Models". In: *Journal of Computational Neuroscience* 7.2, pp. 149–171.

Weber, Alison I. and Jonathan W. Pillow (2017). "Capturing the dynamical repertoire of single neurons with generalized linear models". In: *Neural Computation*.

Wedderburn, R. W. M. (1976). "On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models". In: *Biometrika* 63.1, p. 27.

Winfree, Arthur T. (1967). "Biological rhythms and the behavior of populations of coupled oscillators". In: *Journal of Theoretical Biology* 16.1, pp. 15–42.

— (2001). *The Geometry of Biological Time*. Springer.