Inference from stochastic processes with application to birdsongs and biomedicine

PhD Thesis

Mareile Große Ruse

Department of Mathematical Sciences
University of Copenhagen

This thesis has been submitted to the
PhD School of the Faculty of Science,
University of Copenhagen

Author:

Mareile Große Ruse

Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
DK-2100 Copenhagen Ø
mareile@math.ku.dk / mgrosseruse@gmail.com

The PhD thesis was submitted to the PhD School of the Faculty of Science, University of Copenhagen, on 31st of December, 2017.

Supervisor:

Susanne Ditlevsen
Department of Mathematical Sciences
University of Copenhagen
DK-2100 Copenhagen Ø

Co-Supervisor:

Maria Sandsten
Centre for Mathematical Sciences
Lund University
SE-22100 Lund

Co-Supervisor:

Adeline Samson
Université Grenoble Alpes
Laboratoire Jean Kuntzmann
F-38000 Grenoble

Assessment committee:

Helle Sørensen (chair)
Department of Mathematical Sciences
University of Copenhagen
DK-2100 Copenhagen Ø

Sophie Donnet
UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay
FR-75005 Paris

Dan Stowell
Machine Listening Laboratory, Centre for Digital Music
Queen Mary University of London
GB-E14NS London

Acknowledgments

Towards the end there is a thesis. A "partial fulfillment" of the formal requirements for obtaining the degree Doctor of Philosophy. It certainly reflects one facet of my scientific journey, but behind the scenes, it is also partially representative for a personal journey. I would like to direct special thanks to people who have accompanied me on this at times somewhat rocky, but certainly exciting road.

First and foremost, I would like to thank Susanne Ditlevsen for invaluable, constant scientific guidance, and for continuously inspiring me with your positive, energetic personality and scientific curiosity. Adeline Samson, thank you for interesting discussions on scientific and personal matters, and for making my research visits at the LJK in Grenoble so pleasant. I want to express special gratitude to Maria Sandsten for being so welcoming from my very first day in Sweden, for your scientific advice and open ear throughout my research time in Lund and Copenhagen.

I am deeply grateful to the Division of Mathematical Statistics at Lund University for having been incredibly welcoming even during my research time at the University of Copenhagen. Andreas, for allowing me to occupy office space. James, Mona, Lise-Lotte, for putting this into reality. Maria L. and James, for making me smile every morning when I entered the lunchroom.

My warmest thanks go to my friends and family who always have been there for me. In particular, I would like to mention my friends from beachvolleyball, for uncountable energizing hours in the sand. Unn, Mikaela, Rachele, for sharing precious time at and beyond work, and Bjarke, for your positive energy and patience during the finalization. I feel special gratitude towards Timo, for your truly invaluable year-long support and encouragement. Last but certainly not least, I want to mention Alide and my parents. Thank you for your unconditional love and unquestioned support. Throughout, you have been the solid rock and home in my life.

Mareile Große Ruse

Abstract

This thesis contains three contributions on inference from stochastic processes. The first article, which originates from research conducted at Lund University, has a signal processing spirit. The stochastic processes are bird songs and we approach inference from their time-frequency domain representation. We suggest an algorithm for the automated structural analysis of bird songs, which is particularly suitable for noisy recordings and complex song structures. The novel way of assessing similarity between syllables is based on a particular feature representation, which is derived from the syllables' Ambiguity spectra. The other two articles, which present research carried out at the University of Copenhagen, base inference on time-domain representations of stochastic processes. Focus lies on deterministic and stochastic differential equations models with random effects and applications to biomedical data. In Paper II we employ a delay differential equations model with random effects to gain hitherto unknown insights on the initial distribution and metabolism of selenomethionine in the human body. Paper III considers inference for multivariate stochastic differential mixed effects models and has a stronger theoretical spirit. By allowing the inclusion of subject-specific covariate information in the drift, we leave the setting of identically distributed processes. We derive the Maximum-Likelihood estimator from the continuous-time likelihood, prove its consistency and asymptotic normality, and study the bias arising from time-discretization. The method is applied to the statistical analysis of a data set containing EEG recordings from epileptic patients.

Dansk resumé

Denne afhandling består af tre bidrag, der omhandler inferens i stokastiske processer. Den første artikel, der stammer fra forskning udført på Lund Universitet, er indenfor digital signalbehandling. De stokastiske processer er fuglesang, og vores tilgang er baseret på analyser i tids- og frekvensdomænet. Vi foreslår en algoritme til automatisk analyse af strukturen i fuglesang, som er særligt velegnet til støjfyldte optagelser og komplekse sangstrukturer. Den nye måde at kvantificere similariteten mellem stavelser er baseret på en bestemt repræsentation af egenskaber, der er afledt af stavelsernes Ambiguity Spectra. De to andre artikler, der stammer fra forskning udført på Københavns Universitet, baserer den statistisk inferens på stokastiske processer repræsenteret i tidsdomænet. Fokus ligger på deterministiske og stokastiske differentialligningsmodeller med tilfældige effekter og anvendelser på biomedicinske data. I artikel II anvender vi en Delay Differential Equation model med tilfældige effekter for at få indsigt i metabolismen under de første to timer efter injektion af selenomethionin i menneskekroppen. Artikel III omhandler inferens for multivariate stokastiske differentialligningsmodeller med tilfældige effekter, og har en stærkere teoretisk fundering. Ved at tillade subjektspecifikke kovariater i driften af diffusionsmodellen, har vi ikke længere gentagne målinger af identisk fordelte processer. Vi udleder Maximum Likelihood estimatoren fra den kontinuerte-tids likelihood, beviser konsistens og asymptotisk normalitet, og studerer den bias, der opstår fra tidsdiskretiseringen. Metoden anvendes til den statistiske analyse af et datasæt bestående af EEG-målinger fra patienter med epilepsi.

List of Published and Submitted Work

Included:

Große Ruse, M., Samson, A. and Ditlevsen, S. (2017)
Inference for biomedical data using diffusion models with covariates and mixed effects. *Submitted.*

Große Ruse, M., Hasselquist, D., Hansson, B., Tarka, M., Sandsten, M. (2016)
Automated analysis of song structure in complex bird songs. *Animal Behaviour*, **112**, 39-51.

Große Ruse, M., Søndergaard, L.R., Ditlevsen, S., Damgaard, M., Fuglsang, S., Ottesen, J.T., Madsen, J.L. (2015)
Absorption and initial metabolism of ^{75}Se-L-selenomethionine: a kinetic model based on dynamic scintigraphic data. *British Journal of Nutrition*, **114**, 1718-1723.

Not included:

Große Ruse, M., Ritz, C., Hothorn, L.A. (2017).
Simultaneous inference of a binary composite endpoint and its components. *Journal of Biopharmaceutical Statistics*, **27** (1), 56-69.

Sandsten, M., **Große Ruse, M.**, Jönsson, M. (2016)
Robust feature representation for classification of bird song syllables. *EURASIP Journal on Advances in Signal Processing* **2016** (1), 1-16.

Stucke, D., **Große Ruse, M.**, Lebelt, D. (2015)
Measuring heart rate variability in horses to investigate the autonomic nervous system activity - Pros and cons of different methods. *Applied Animal Behaviour Science*, **166**, 1-10.

Stucke, D., Minero, M., Dalla Costa, E., **Große Ruse, M.**, Langbein, J., Hall, S., Lebelt, D. (2015)
Praxistaugliche Schmerzindikatoren beim Pferd. *DVG Service GmbH Verlag, GieÃ§en*, ISBN 978-3-86345-243-8, pp. 157-168.

Stucke, D., **Große Ruse, M.**, Langbein, J., Lebelt, D. (2014).
Die Bedeutung der Herzfrequenzvariabilitätsanalyse als Schmerzindikator bei Pferden - The validity of heart rate variability analysis to identify pain in horses. *KTBL-Schrift 504, Aktuelle Arbeiten zur artgemäßen Tierhaltung 2014, Kuratorium für Technik und Bauwesen in der Landwirtschaft e.V. (KTBL), Darmstadt.*

Stucke, D., Hall, S., **Große Ruse, M.**, Lebelt, D. (2013).
Untersuchung zur Quantifizierung von Schmerzen bei Pferden - Investigation to quantify pain in horses. *KTBL-Schrift 503, S. 116-125, Aktuelle Arbeiten zur artgemäßen Tierhaltung 2013, Kuratorium für Technik und Bauwesen in der Landwirtschaft e.V. (KTBL), Darmstadt, ISBN 978-3-941583-87-0.*

Summary

This thesis contains three research papers (two published, one submitted for publication), which are summarized below and attached at the end of the thesis. Chapter 1 motivates the three research projects and summarizes our main contribution. Chapter 2 reviews essentials of frequency and time-frequency analysis of stochastic signals, which constitute useful background information for methods used in the first paper. While the knowledge of facts presented in Chapter 2 is not required for Paper 1, they do enhance the understanding of a reader who is not too familiar with frequency- and time-frequency domain analyses. The first part of Chapter 3 motivates the inclusion of various sources of noise in differential equation models, and discusses the challenges for parameter estimation that come with the resulting model complexity. We provide additional material on Paper II, including a motivation for the kinetic model from the article, a description of parameter estimation with Monolix (2016), and individual model fits. The last section in Chapter 3 contains supplementary material on the third research article. We embed the stochastic differential mixed effects model (SDMEM) into a rigorous formal setting, elaborate on regularity assumptions and provide detailed proofs on asymptotic properties of the MLE.

I - Automated analysis of song structure in complex birdsongs

In the first project, we introduce an algorithm for the automated detection and clustering of bird song syllables. The three-step method is fully self-contained and thereby enables a time-efficient analysis of birdsongs without observer bias. Objectiveness is particularly important to ensure comparability of results between different studies. A compelling feature of the method is its trade-off between noise robustness on the one hand, and detail focus on the other. This makes it particularly suitable for noisy field recordings of complex songs. The algorithm's performance is demonstrated on field recordings of Great Reed Warbler songs and benchmarked against human expert analysis and other established methods.

II - Absorption and initial metabolism of ^{75}Se-L-selenomethionine: a kinetic model based on dynamic scintigraphic data

Paper II is concerned with statistical inference in pharmacokinetics. It provides a thorough statistical analysis of high-frequency imaging data in order to gain insights on absorption and metabolism of Selenomethionine during the first two hours after oral intake. Inference is based on a three-layer kinetic model. The first layer describes the underlying dynamics by means of a multidimensional delay differential equation, the second accounts for bias arising from overlapping tissues in the images, and the third layer incorporates measurement uncertainties. Subject-specific variations in both dynamics and measurement variances are captured through the inclusion of random effects. The model that achieved the best fit to the data was considerably simpler than compartmental models on Selenium metabolism suggested by literature. This can partly be explained by the different nature of data. Previous studies merely focused on long-term dynamics, whereas the current study aimed at gaining insights on the initial dynamics based on high-frequency measurements. Due to the high temporal resolution, a hitherto unknown plasma plateau could be discovered, and explained by the model. The obtained rate and delay parameters can be useful for assessing intestinal absorption capacity or liver function in patients.

III - Inference for biomedical data using diffusion models with covariates and mixed effects

In Paper III we consider more generally the statistical inference for longitudinal biomedical data. Some features are characteristic for many of these data sets, (i) a high sampling frequency, which makes them very suitable for continuous-time modeling approaches, (ii) inter-subject variability that can be explained by covariates, (iii) complex dynamics of an often multivariate data generating process, and (iv) a substantial amount of unexplained variation (noise), which calls for models that incorporate different types of stochasticity. These properties lead naturally to stochastic differential mixed effects models (SDMEMs), which allow for non-linear dynamics, multivariate states and the inclusion of covariates. We investigate likelihood-based parameter inference for this model class. In particular, we study the asymptotic behavior of the continuous-time Maximum Likelihood estimator theoretically and in simulations, and analyze the bias caused by its time discretization. Hypothesis testing is discussed as well, and explored in simulations. We finally apply the model framework to the statistical analysis of EEG data from epileptic patients.

Contents

Chapter 1

Introduction

1.1 Bird song - What is it and why do people study it?

Bird songs are among the most fascinating sounds we hear in everyday life, and are often associated with the anticipation of an upcoming spring. Even urban areas can be scenes for a variety of avian vocalizations, e.g., the curring of pigeons (*Columbidae*) or the song of a house sparrow (*Passer domesticus*) sitting on the balcony or in the garden. These vocalizations are, however, profoundly different, even for the ornithologically untrained ear. The pigeon's sounds are non-melodic, of very simple structure and can hardly be called a *song*, whereas the sparrow pleases our ears with much more melodic, song-like performances. The trained ear may even detect differences between the songs of birds from the same species. But when exactly can a vocalization be called a *song*? Is the curring of the pigeon a song, only a very simplistic one? Surprisingly enough, despite a several decades long interest of researchers in avian singing, there exists, really, no clear definition of the term *song*. However, a common understanding is to think about a song as a rather complex and often melodious sequence of $50 - 300ms$ long elements, the so-called syllables (Catchpole and Slater, 2008). With this definition, pigeon sounds are not songs, but instead considered as simple *calls*. Whereas the sparrow, the common blackbird (*Turdus merula*), the nightingale (*Luscinia megarhynchos*), or the warblers, are examples of *songbirds* (*Passeri*), a class of birds comprising several thousands of species, which are able to produce complex vocalizations such as songs.

Three scientific reasons for the interest in bird songs

The song's complexity is understood as the number of different syllable types that occur in the song, which is also referred to as the syllable *repertoire*. In the first part of this thesis, we investigate

the syllable repertoire of the Great Reed Warbler (*Acrocephalus arundinaceus*, GRW for short), one of which is shown in the left hand side of Figure 1.1. The right hand side illustrates a part of its song with the visually clearly separated syllables. The two somewhat broader syllables in the middle, e.g., are realizations of the same syllable type, and so are the following four syllables of high amplitude.

Beyond the magic that humans associate with bird songs, there are (at least) three more key reasons why researchers from all over the world engaged in investigations of complex bird vocalizations. The first one goes almost 150 years back to Charles Darwin (1871), who postulated that the *"sounds uttered by birds offer in several respects the nearest analogy to language"*. In fact, modern research confirms that he was on the right track, and this has wide-range implications. Bird song is in structure indeed very similar to human language. Both consist of syllables which are combined to larger entities (*motives* or *strophes* for birds, *words* for humans), and which themselves build a bird's song or a sentence. Most important, however, is that the avian language is *learned* by the young bird through imitation, just like humans learn to speak by imitating their parents. Current studies have shown that this comes with considerable behavioral, neural and genetic parallels between bird and humans (Berwick et al., 2011; Brainard and Doupe, 2002, 2013; Elemans et al., 2015). Thus, the songbird can be used as a model system for how humans learn and vocalize language, or more generally, how they learn and execute any sequence of actions. The rich existing knowledge on the neural mechanisms which drive learning and vocalization in birds can thus be used to gain a better understanding on the corresponding mechanisms in the human brain.

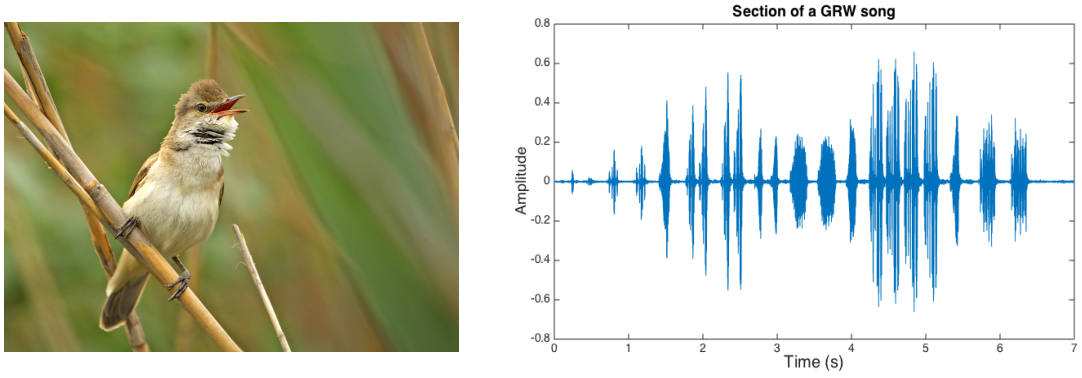


Figure 1.1: The Great Reed Warbler (left) got its name from its preferred breeding habitat, the reed beds. A section of its song is displayed in the right figure.

Another aspect that made researchers focus their attention on the study of bird songs is species

recognition (Miller, 1997; Somervuo et al., 2006; Potamitis et al., 2014). This is useful for an automated and objective assessment and continued monitoring of bird diversity (e.g., to examine the quality of an ecosystem). Songs can differ considerably between species and are therefore a good characterizing feature. Moreover, audio-based recognition techniques are especially useful in densely vegetated areas, where difficult visual conditions make it impossible to spot and visually identify the singing birds.

Thirdly, and at least for this thesis most importantly, there is the ecological role of the song for the bird itself, which has since long spurred ornithologists' interest and which builds the motivational ground for song analysis in this thesis. During mating season, male birds spend remarkable energy on broadcasting their songs (Miller and Kroodsma, 1996; Catchpole and Slater, 2008). The purpose is to attract female mating partners and to defend territory. Certainly, a melodic, complex song appears more appealing to the human ear than a repetitive sequence of a few simple calls. That this impression can to some extend be transferred to the perceptions of female birds (even though with a possibly different motivation) was indicated in several studies. Researchers could show that the complexity of a bird song is positively correlated with the vocalist's reproduction success (harem size, number of offspring and the offspring's survival probability) and the quality of its territory (Catchpole, 1986; Hasselquist, 1998; Nowicki et al., 2000). Thus, a bird's syllable repertoire size is a helpful metric in avian studies and an objective way for the qualitative and quantitative investigation of syllable repertoires is sought for.

Research contribution

The underlying ecological research questions for the first thesis project were investigations of how song repertoires of GRWs differ (i) within one individual over time, (ii) between individuals of the same population, and (iii) between individuals of different populations. The data, comprising song recordings of GRWs from more than 30 years, has been collected by our collaborators Dennis Hasselquist, Bengt Hansson and Maja Tarka. Despite the existence of various bird analysis tools, they lacked one that had the resolution and noise suppression required to reliably, and objectively, analyse noisy recordings of complex GRW songs. This initiated the research on a suitable automated analysis tool, the result of which is presented in Paper I.

In this article, we suggest an algorithm which enables an objective and automated investigation of a bird's syllable repertoire. The algorithm comprises three steps, (i) the syllable extraction, (ii) their representation and comparison based on selected features, and (iii) the clustering of syllables. The extraction step was already presented in a preliminary work (Hansson-Sandsten et al., 2011),

which was also the first literature source to investigate an SVD-based feature representation of syllables in the ambiguity domain. We build upon their ideas in order to obtain a fully self-contained clustering method with benchmarked and validated performance[1]. More specifically, we keep the ambiguity feature space, but use a different similarity measure, and we provide a profound investigation of the algorithm's performance by benchmarking it against human expert analysis and other state-of-the-art methods. We append a clustering process and compare the results to human expert analysis, based on a data set containing more than 400 syllables. To facilitate straightforward application for the practitioner, we provide ready-to-use code upon request.

The available data consists of field recordings of GRW songs, which are typically very noisy. This may be due to vocalizations of neighboring birds or other animals close by, or to unfortunate weather conditions such as wind or rain. On the other hand, the songs of GRWs are fairly complex, such that it is not immediate to determine, whether two syllables are merely noisy realizations of the same syllable type, or if their differences are systematic and the syllables are examples of two different types. A good feature representation has to manage the balance of keeping characterizing details, while disregarding unimportant syllable information. It is shown that the combination of (a) feature selection, which is based on a noise-robust and low-dimensional syllable representation in the Ambiguity domain, and (b) similarity assessment, which simultaneously investigates syllable resemblance in time- and frequency dimension, succeeds in exactly that.

Possible extensions

The algorithm was only applied to song recordings from the GRW. It would be interesting to investigate its performance for song data of other bird species. Even though our method is fully self-contained, it does require the practitioner to calibrate a few parameters prior to analysis (such as the number and the specific choice of tapers for the multitaper spectrogram). It is of interest to see whether parameters that were chosen optimally for the GRW have to be adjusted to also ensure a high performance for songs from other species.

One popular application of song analysis tools which we have not touched upon in our work is bird species recognition. Song-based species classification is important in monitoring the state of ecosystems, but it has also gained much popularity in leisure context as a result of the increasing interest of people in bird watching. Literature has suggested a compelling variety of possible

[1] A further exploration of different syllable representations, parameter settings and similarity measures was conducted in our later work (Sandsten et al., 2016).

feature representations, but features derived from the Ambiguity spectrum of a syllable have not been considered in this application area.

1.2 Inference for biomedical data

Typical characteristics of data in biomedical applications

Rapidly advancing technology has rendered massive data collection and storage inexpensive and feasible for institutions as well as individuals. In healthcare, hospitals and medical clinics make increasingly use of electronic devices, such as wearable sensors, smartphones, or dynamic imaging apparatus to monitor health-related variables in patients (Xie et al., 2017). Examples of such intensive longitudinal data (ILD) are the passive tracking of a patient's blood pressure or glucose level (Walls and Schafer, 2006), dynamic gamma imaging in pharmacokinetic studies (Paper II), or long-term measurements of electrical activity in the brain (electroencephalography, EEG) on epileptic patients (Paper III).
Since devices can easily monitor several variables simultaneously, and measure them at high frequency (compared to the typical time scale of the observed system), ILD naturally lend themselves for multidimensional, continuous-time modeling. Differential equation models have become a popular tool for describing the continuous-time dynamics of a multidimensional system. As Walls and Schafer (2006) point out, however, those dynamics are often complex and characterized by a considerable "*variety of individual trajectories*". This heterogeneity between subjects can typically not solely be explained by covariate information. Therefore, the inclusion of random effects which account for unexplained inter-subject variability becomes a necessity (Timms et al., 2014; Dziak et al., 2015).

Drawbacks of ordinary differential equation models with random effects

Ordinary differential equations (ODEs) with random effects have frequently been applied to model biomedical data, (Ribba et al., 2014; Jarne et al., 2017). Their formulation is intuitive, the random effects capture inter-individual deviations from the population dynamics, and today's computational power renders parameter estimation feasible. Well-known applications of this model framework are pharmacokinetic compartment models (Tornøe et al., 2004; Lavielle, 2014), which are used to describe the flow of a substance between multiple spatially separated entities (different organs in the human body). A drawback of ODE-based models is the deterministic nature of ODEs themselves, which mirrors the implicit assumption that the system's dynamics is entirely captured

by the model equations. Biological systems are, however, incredibly complex. Their variability is driven by the interplay of numerous internal (genetic variations, metabolic fluctuations, etc.) and external factors (stress factors, room temperature, time of day, etc.). The bulk of them can not be measured directly, or can not be included in the model, because it would prohibitively scale up the model's complexity. However, ignoring those inadequacies or uncertainties in the model structure lead, if they are substantial, to biased estimates and false inference (Donnet et al., 2010).

Challenges of stochastic differential equation models with random effects

In those cases one can achieve a more robust estimation by replacing ODEs with stochastic differential equations (SDEs), which account for unexplained system-intern fluctuations through the incorporation of a stochastic term (Møller et al., 2010; Leander et al., 2014). As compared to the vast amount of literature that exists on parameter estimation for standard SDE models, statistical inference for SDEs with random effects (so-called stochastic differential mixed effects models, SDMEMs) has caught researchers' attention only about a decade ago. Since their introduction, research on SDMEMs has quickly gained momentum. This is particularly due to two factors.
The first one is application-driven. Especially in biomedicine, SDMEMs have and will continue to have wide applicability (with examples mentioned above). They allow modeling of complex longitudinal data and, at the same time, enable robust statistical inference. The other factor triggers the theoretical and computational enthusiasts in the statistical community. Also in statistics there is nothing like a free lunch. The certainly powerful merging of random effects and SDEs into one single model comes with a considerable challenge for inference based on the data likelihood: its intractability. This now has two sources. First of all, the likelihood for the SDE part, assuming the random effects were observed, is typically unknown. But even if it was known, this likelihood has to be marginalized over the distribution of the random effects, because the random effects are practically not observed. The marginalization is an (often multidimensional) integral, which rarely has a closed-form solution. This makes explicit likelihood inference impossible and leaves many research opportunities for finding well-performing numerical or analytical approximation techniques. In fact, numerous ways of tackling this challenge have been explored. In section 3.2 we provide a short literature overview of the main approaches.

Research contribution

A substantial amount of fluctuation in biomedical data can not explicitly be explained by the model. This calls for model frameworks which capture unexplained variability by the inclusion of

measurement noise, intrinsic stochasticity and/or unexplained subject-specific fluctuations. However, likelihood-based parameter inference in differential equation models that account for all three noise sources is intractable, and one has to revert to numerical or model approximations. If some of the noise sources are negligible as compared to others, models with fewer noise terms are more tractable and can often still facilitate reliable inference.

In a pharmacological study presented in Paper II we trade model complexity against tractability by reverting to ODE-based models. We provide a comprehensive statistical analysis of data on the initial metabolism of ^{75}Se-L-selenomethionine (SeMet) in humans. Unexplained data variability is captured by the inclusion of measurement noise and random effects, where the optimal types of measurement noise (additive versus multiplicative) and distributions of random effects are carefully explored and selected. We also investigated various models that were previously suggested in literature, most of which were considerably more complex. However, model selection criteria and the spirit of Occam's Razor let us to conclude that a fairly parsimonious, and at the same time intuitive, model yields the best fit for the data at hand. This resulted in new insights on the pharmacokinetic properties of SeMet in the body within the first $2h$ after administration.

In the third project we pursued an alternative way of model approximation, which is motivated by the high-frequency nature of many biomedical data sets. In the framework of multidimensional SDMEMs with covariates, we base parameter inference on the continuous-time likelihood. In contrast to transition densities, which are mostly unknown for SDE models, the continuous-time likelihood is readily available via absolute continuity of measures in the space of continuous functions. The asymptotic properties of the Maximum Likelihood estimator (MLE) (when the number of subjects grows to infinity), and the bias that is introduced when the likelihood is discretized in time, are studied theoretically and in simulations. The promising performance motivates the application of this method to real data, where we provide a statistical investigation of EEG data from epileptic patients using SDMEMs with covariates.

Chapter 2

Time-frequency analysis of signals

To assess the complexity of a bird song, its building blocks, the syllables, have to be compared and classified. To this end, every syllable is represented by a collection of characterizing *features*, which could be its duration, its frequency range, or more involved attributes (see Paper I). Syllable comparison (and clustering) is then performed based on the selected features. The quality of any clustering or classification algorithm therefore depends crucially on how one chooses to represent the syllables. A good feature set separates the signal from the noise in an efficient manner, highlighting the most important distinguishing characteristics and disregarding irrelevant information.

Features should include information on a syllable's frequency content

A song syllable, being an auditive signal, is composed of a variety of waveforms at different frequencies, which are produced in the the avian vocal organ (syrinx) by pressing air from the lungs over vibrating membranes (Catchpole and Slater, 2008). The waveforms are therefore the elementary units in a syllable. This indicates that a syllable is well characterized by its frequency content and that a good feature set for song syllables should include some form of frequency information. The frequency content, or spectrum, quantifies how much of a syllable's variations is due to oscillations at particular frequencies. Frequencies are usually measured in Hertz (Hz), the number of cycles per second (s), and for a bird song they typically lie between 100 to 500 Hz (though the range differs substantially across bird species and is usually more narrow for a specific species).

Frequencies do not carry temporal information

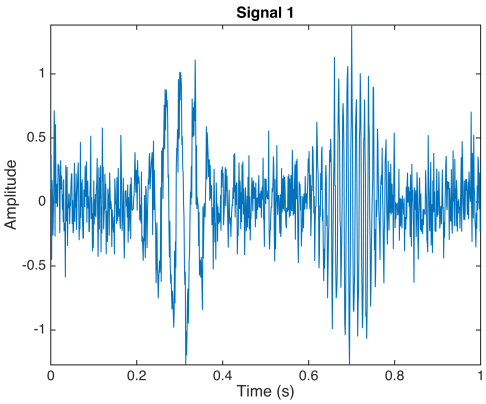
The standard way to represent an auditive signal is via its amplitude over time (cf. the upper two panels in Figure 2.1). To analyze a syllable's spectrum, it has to be Fourier transformed into the

frequency domain. Since this operation is an integration of the (exponentially weighted) syllable signal over the whole time course, the frequency content is an inherently *global* property and does not contain temporal information. In other words, it does not reveal *when* in time a syllable contains a significant contribution at a particular frequency. This is illustrated in Figure 2.1. The stochastic process displayed in the upper left panel consists of two Gaussian-envelope-type components with frequencies 30 Hz and 100 Hz (sampled at frequency $F_s = 1000$ and observed during a time interval of $1s$), respectively. The upper right panel shows a signal with the same Gaussian components, but in reversed order. The estimated spectra (obtained by a Welch periodogram, see section 2.1) shown in the lower two panels do not reveal the different occurrence times of these two components, because time information has been lost by applying the Fourier transform.

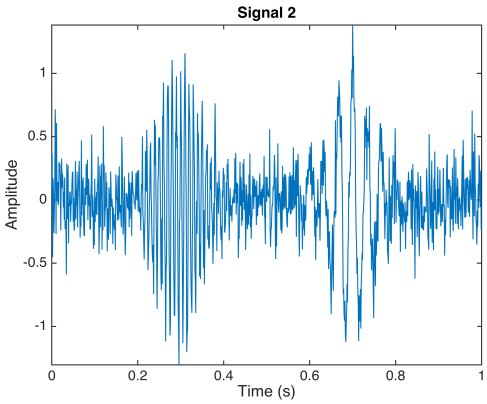
The frequency content of a song syllable is time-dependent

The inherent global property of a standard frequency is not an issue when the underlying signal is *stationary*, i.e., when its frequency content does not change over time. However, bird songs are not stationary, a time-varying frequency content is very common. As one may imagine, a male bird's song that exclusively contains syllables with time-constant spectral content has a less advertising character than a song with more temporal variations. Where courtship and mating success rely heavily on auditive advertisement, an inspiring complex song is therefore an advantage in the competition for potential female mating partners. To capture the time-dependency of a syllable's spectral content, one therefore migrates from frequency representations to *time-frequency* representations. These are two-dimensional transformations of the signal and can be interpreted as time-dependent Fourier transforms. To set notations and render the concept of time-frequency distributions more accessible, we start with a short review on spectral analysis of stationary signals in section 2.1, before diving more deeply into the field of time-frequency representations for non-stationary signals in section 2.2.

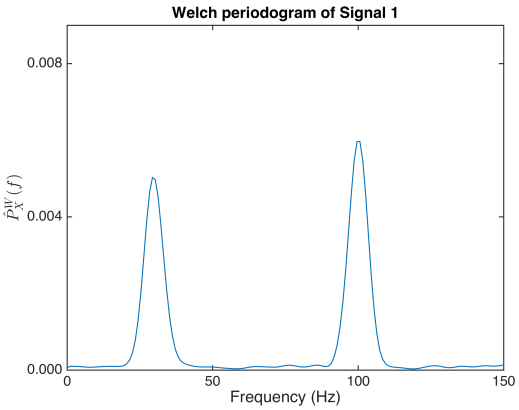
In this thesis, we treat a signal such as a bird song as a *stochastic process* $X = (X_t)_{t\in\mathbb{T}}$, i.e., X_t is a random variable and $\mathbb{T}$ is an index set, usually representing time, and is commonly chosen as the whole real line, a finite interval $[0, T]$, or a discrete set of measurement time points $\{t_1, \ldots, t_n\}$. The process X is called (wide-sense / second-order) *stationary*, if its expectation $\mathbb{E}(X_t)$ does not change as a function of $t \in \mathbb{T}$, and if the autocorrelation function (ACF) of X, $r_X(t, s) = \mathbb{E}\left([X_t - \mathbb{E}(X_t)]\,[X_s - \mathbb{E}(X_s)]'\right)$ only depends on the time lag $t - s$, but not on the temporal location, i.e., $r_X(t, s) = r_X(t - s)$. We assume that $\mathbb{E}(X_t) = 0$ whenever X is stationary.



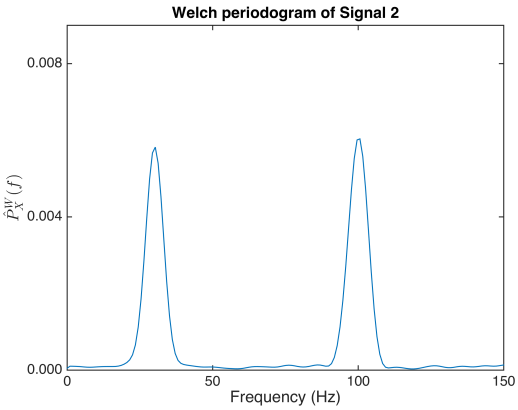
(a) Stochastic process with two Gaussian-envelope-type contributions, at 30 Hz centered at time $t = 0.3s$ and at 100 Hz centered at $t = 0.7s$.



(b) Stochastic process as in 2.1(a), but with frequency components in reversed order.



(c) Welch periodogram of Signal 1



(d) Welch periodogram of Signal 2

Figure 2.1: Two stochastic signals in their time representation in the upper panels and their corresponding estimated spectra (Welch periodogram with 50% overlap and Hanning window of length $0.25s$) in the lower panels. The signals have the same frequency content, but at different times. This is clearly visible in the time representation, but the temporal information is completely lost in the spectral representation.

2.1 Frequency analysis of stationary signals

The power spectrum of a stationary stochastic process

For a stationary stochastic process $X = (X_t)_{t\in\mathbb{T}}$, the variance $r_X(0) = \mathbb{E}(|X_t|^2)$ is also called the *power* of X. This is the statistical variation of the process at any given point in time. If the covariance function of X is integrable, it can be obtained via the integration $r_X(\tau) = \int_{-\infty}^{\infty} P_X(f)\mathrm{e}^{i2\pi f\tau}df$, where

$$P_X(f) = \mathcal{F}_{\tau\to f}\left[r_X(\tau)\right] = \int_{-\infty}^{\infty} r_X(\tau)\mathrm{e}^{-i2\pi f\tau}d\tau$$

is the Fourier transform of r_X. Since the power of X can therefore be written as $r_X(0) = \int_{-\infty}^{\infty} P_X(f)df$, the function P_X basically describes how power of X is distributed over different frequencies. Therefore, P_X is called the *power spectral density* (PSD) of r_X (or X). Sometimes, P_X is also simply called the *spectrum*. This is the quantity the frequency-domain analysis of a signal X is interested in and aims to estimate.

The periodogram and a heuristic motivation

The periodogram is the simplest estimator of the PSD and the basis for more advanced spectral estimators. It relies on finitely many discrete-time observations of a single realization of the signal X. More specifically, we assume to measure X during some time interval $[0, T]$ at a sampling frequency $F_s = 1/\Delta t$, where Δt is a positive time interval. The measurement time points are $t_k = k \cdot \Delta t$ for $k = 1, \ldots, n$, with $n = T/\Delta t$ as the total number of samples. Based on the obtained data $x_k := X_{t_k}, k = 1, \ldots, n$, the periodogram is defined as

$$\hat{P}_X^{\mathrm{per}}(f) = \frac{1}{F_s n}\left|\sum_{k=1}^{n} x_k \mathrm{e}^{-i2\pi\frac{f}{F_s}k}\right|^2, \qquad -\frac{F_s}{2} < f \leq \frac{F_s}{2}. \tag{2.1}$$

The frequency resolution of the periodogram, that is, the minimum distance between two frequencies required in order to be resolved by the periodogram, is $\frac{1}{T} = \frac{1}{n\Delta t}$. In particular, we can only increase the resolution by measuring for a longer time (increasing T), not by sampling more frequently. To avoid aliasing (introducing false frequency content to lower frequencies), the sampling frequency must be larger than twice the maximum frequency that is present in the signal (Lindgren et al., 2013).

The connection of the PSD estimate (2.1) to the PSD itself might not be immediate. Therefore, and because of the periodogram's basic importance, we give a heuristic motivation of the periodogram instead of just leaving the reader with its definition. To start with, note that for large T (and

sufficiently quickly decaying r_X) we can approximate $P_X(f)$ reasonably well by

$$P_X(f) = \int_{-\infty}^{\infty} r_X(\tau)\mathrm{e}^{-i2\pi f\tau}d\tau \approx \int_{-T}^{T} r_X(\tau)\mathrm{e}^{-i2\pi f\tau}d\tau.$$

This can be rewritten as

$$\begin{aligned}
&= \frac{1}{T}\int_{-T}^{T} r_X(\tau)\left(\int_{\tau-T}^{\tau} 1\,dt\right)\mathrm{e}^{-i2\pi f\tau}d\tau = \frac{1}{T}\int_{-T}^{T}\int_{\tau-T}^{\tau} r_X(\tau)\mathrm{e}^{-i2\pi f\tau}dtd\tau \\
&= \frac{1}{T}\int_{0}^{T}\int_{t-T}^{t} r_X(\tau)\mathrm{e}^{-i2\pi f\tau}d\tau dt = \frac{1}{T}\int_{0}^{T}\int_{0}^{T} r_X(t-s)\mathrm{e}^{-i2\pi f(t-s)}dsdt.
\end{aligned}$$

We estimate $r_X(t-s) = \mathbb{E}(X_tX'_s)$ by $X_tX'_s$ and obtain the approximation

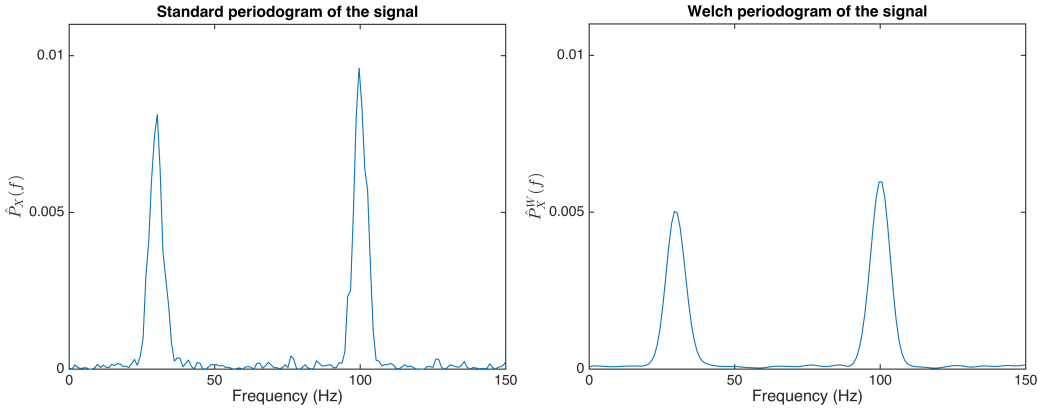
$$\begin{aligned}
&\approx \frac{1}{T}\int_{0}^{T}\int_{0}^{T} X_tX'_s\mathrm{e}^{-i2\pi f(t-s)}dsdt \\
&= \frac{1}{T}\left|\int_{0}^{T} X_t\mathrm{e}^{-i2\pi ft}dt\right|^2 =: \hat{P}_X(f).
\end{aligned} \tag{2.2}$$

$\hat{P}_X(f)$ is a continuous-time estimate of the PSD, based on the observation of a single trajectory during the time interval $[0,T]$. The periodogram is simply the time-discretization of $\hat{P}_X(f)$ (using $\frac{1}{F_s n} = \frac{1}{T}(\Delta t)^2$).

Improving the periodogram: Welch's method and Multitapers

The periodogram suffers from two drawbacks, *bias*, i.e., $\mathbb{E}(\hat{P}_X^{\mathrm{per}}(f)) \neq P_X(f)$, and a high *variance*. The bias has two sources, leakage and limited resolution, and can be altered by *windowing* (also called *tapering*). The periodogram can be rewritten as $\hat{P}_X^{\mathrm{per}}(f) = \frac{1}{F_s}|\sum_{k=1}^{n} x_k h_k \mathrm{e}^{-i2\pi\frac{f}{F_s}k}|^2$, with $h_k = h(k\Delta t)$ and $h(t) = \frac{1}{\sqrt{TF_s}}\mathbb{1}_{[0,T]}(t)$. This operation of multiplying the data sequence by another sequence $(h_k)_{1\leq k\leq n}$ is called windowing and h is the *window function* (or *taper*). Leakage refers to the fact that power from frequency bands with high signal power leaks into other frequency bands. This can be addressed by replacing the rectangular window with a smoother window function. A common choice is the *Hanning window* $h(t) = 0.5 - 0.5\cos(2\pi t/T)$. The resulting PSD estimate is called *windowed periodogram*. However, the reduced leakage of the windowed periodogram comes at the expense of decreased resolution. Various windows have been suggested in literature that balance this trade-off in different ways. The variance is addressed by *averaging* several periodograms, and two approaches can be distinguished. The *Welch method* relies on cutting the data into M (often overlapping) smaller segments, windowing them (for bias reduction), and calculating windowed periodograms for every data segment. Finally, the obtained M windowed periodograms are averaged. The segmentation, however, comes with a loss in resolution, because

every windowed periodogram is now based on a shorter data sequence. Thus, another trade-off has to be faced, this time between resolution (the longer the segment, the better the resolution) and variance reduction (the shorter the segments, the more segments one obtains and thus the higher is the variance reduction).



(a) Standard periodogram estimate.

(b) Welch periodogram estimate

Figure 2.2: Comparison of spectral estimates for the signal displayed in Figure 2.1(a). The standard periodogram (left) is more wiggly than the estimate obtained by the Welch approach (right). The Welch periodogram was computed with a Hanning window of length $0.25s$ and a 50% overlap.

Like the Welch method, also the *Multitaper (MT) approach* (Thomson, 1982) averages windowed periodograms. The difference is that every windowed periodogram of the MT estimate (i) relies on the entire data sequence, and (ii) has its own taper. As orthogonal tapers give maximal variance reduction, the tapers are designed to be orthogonal while reducing bias. Popular multitapers are the Slepian sequences (Slepian and Pollak, 1961) or Hermite windows.

A more detailed review on the PSD and its non-parametric (via the Welch estimator) and parametric (via autoregressive modeling) estimation can be found in our work Stucke et al. (2015), where we apply spectral analysis to heart rate variability data from horses.

2.2 Time-frequency analysis of non-stationary signals

A song syllable is a *non-stationary* signal, but it can be considered stationary on a short time scale. The time-frequency approach is tailor-made for this type of locally stationary (quasi-stationary) data. It transforms the signal into a two-dimensional representation, with one temporal dimension and the other dimension being related to frequency. Thereby, one obtains a time-dependent representation of a signal's frequency content. In this section, we will cover two types of time-frequency

representations, the spectrogram and the Ambiguity spectrum. The spectrogram is a function of time and frequency, and can be seen as a time-dependent periodogram. Since its introduction to bird songs in the 50's, it has become one of the most widely used tools for bird song analysis (Hasselquist et al., 1996; Tchernichovski et al., 2004; Węgrzyn and Leniowski, 2010). On the contrary, the Ambiguity spectrum, which has its origin and name from a radar context (Woodward, 1953), has first been applied to bird songs only recently (Hansson-Sandsten et al., 2011). Being a function of time- and frequency-*lags*, it has inherently different properties than the spectrogram and proves to be very well-suited for syllable characterizations. Despite being rather different in nature, the spectrogram and the Ambiguity spectrum are closely connected via Fourier transformations.

Spectrogram

The spectrogram is an extension of the periodogram to non-stationary processes and relies on the reasoning that short segments of a non-stationary signal can be considered stationary. In continuous time, it is defined as

$$\hat{P}_X(t,f) = \left| \int_{-\infty}^{\infty} X_s g(t-s) \mathrm{e}^{-i2\pi f s} ds \right|^2 . \tag{2.3}$$

The window function g is centered at 0, and acts as a sliding window, extracting shorter segments

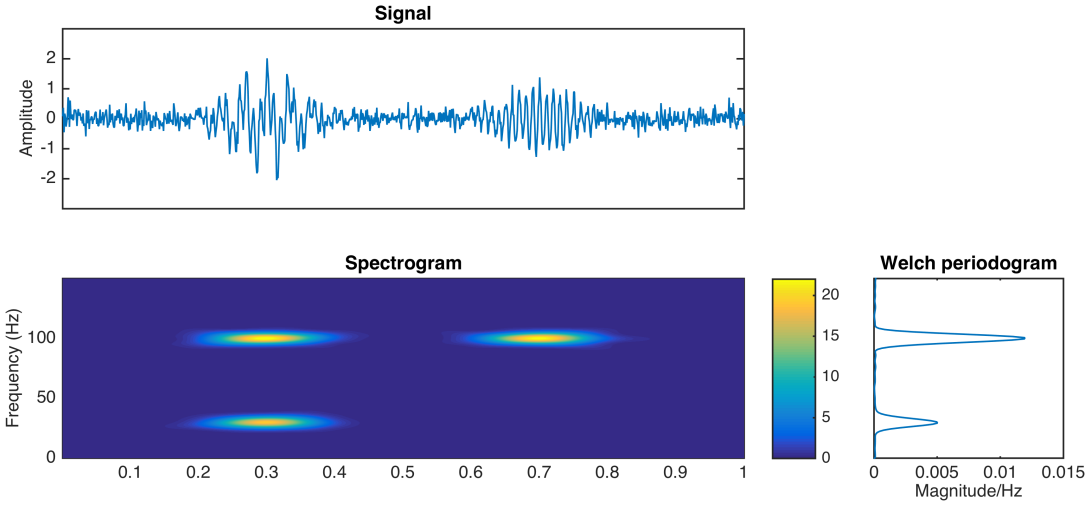


Figure 2.3: Three representations of a signal. The upper panel shows the standard time-domain representation. Its spectrogram (calculated using a Hanning window of length $0.25s$) is displayed in the lower left panel, clearly indicating the location of frequency content in the time dimension. The lower right plot shows the Welch periodogram estimate, which does not reveal temporal information.

from the process that are stationary. In that way, the spectrogram can really be seen as a sliding

(in time t) extension of $\hat{P}_X(f)$ in (2.2) to non-stationary processes. The length of the window g is chosen such that the windowed data segments are stationary. However, a trade-off has to be found between achieving a suitable resolution in frequency (long window), assuring stationarity of the windowed data (short window), and obtaining a sufficient resolution in time (short window). The spectrogram reveals the "local" spectrum of a non-stationary signal at any given time t, and can thereby provide information on temporal changes in a signal's frequency content. This is illustrated in Figure 2.3 by means of a stochastic signal with three time-limited Gaussian-envelope components, two of them (with frequencies 30 and 100 Hz, respectively) are centered at time $0.3s$, and one 100 Hz frequency component is centered at time $0.7s$. The upper panel shows the signal's time-representation, the lower left the spectrogram[1] and the lower right the Welch periodogram. While the spectrogram clearly reveals the three components and their time of occurrence, the Welch estimate is not able to provide temporal information.

Being locally like a periodogram, the (time-discretized) spectrogram suffers from the same drawbacks as the periodogram, which can be addressed in the previously presented manners (tapering, averaging). In Paper I, we work with spectrograms for which bias and variance are reduced by the MT approach. In Figure 2.4 we illustrate the effect of the number of MTs on the resulting MT spectrogram. The second panel displays the spectrogram obtained from using a single Hermite window. The third plot shows the MT spectrogram with two Hermite tapers and the bottom panel illustrates an MT spectrogram with four Hermite tapers. The resolution in time- and frequency domain is much better with fewer tapers. This gain comes, however, at the expense of a larger variance.

Ambiguity spectrum

The Ambiguity spectrum[2] (AS) is at the heart of our chosen feature representation for syllables and is therefore a key ingredient to the automated bird song analysis method. It is defined as the Fourier transform (in t) of the centered autocorrelation,

$$A_X(\nu,\tau) = \int_{-\infty}^{\infty} r_X(t-\tfrac{\tau}{2}, t+\tfrac{\tau}{2}) \mathrm{e}^{-i2\pi\nu t} dt. \tag{2.4}$$

The argument ν is the *Doppler lag* and τ is the *time lag*. Characteristic for the AS is the invariance (of its absolute value) to time- and frequency shifts of the signal. This is illustrated in Figure 2.5, where we consider three Gaussian-envelope-type stochastic signals (first column). The first one

[1] Unless plotted otherwise, the colorbar in this figure is representative for all subsequent spectrogram figures.
[2] The term "Ambiguity" in this context has its justification from a radar context in Woodward (1953).

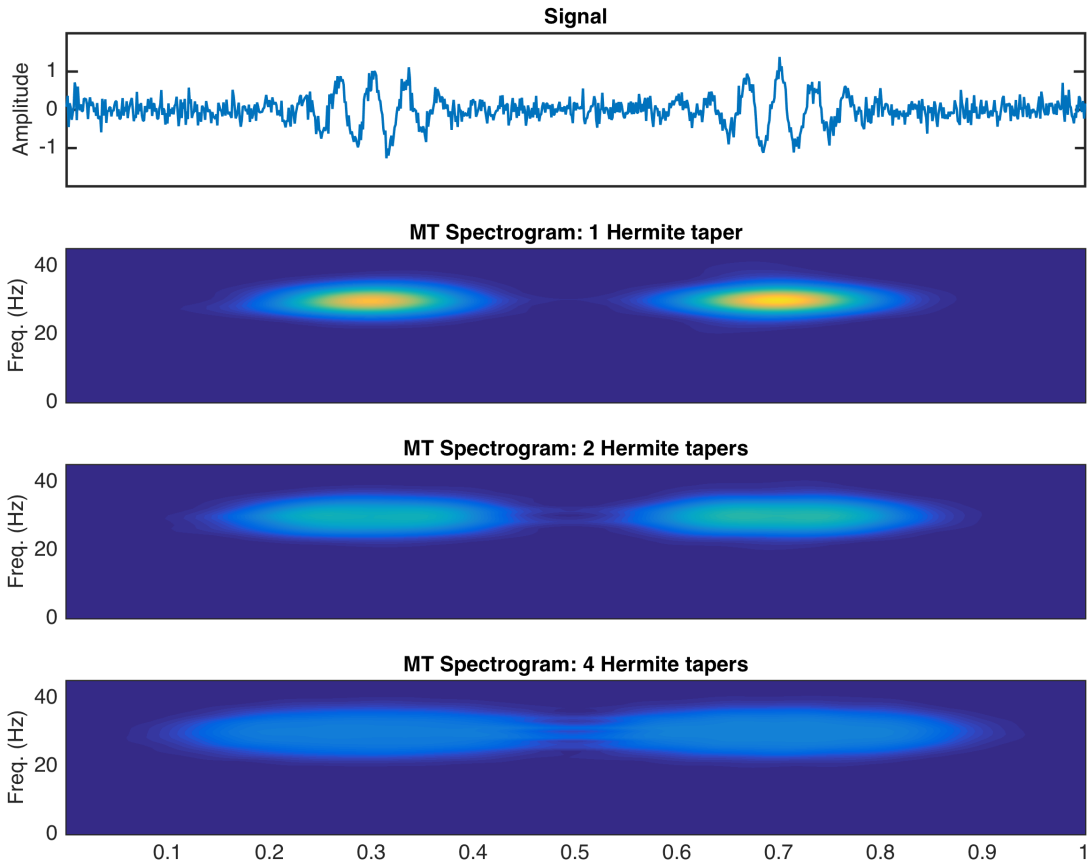


Figure 2.4: Comparison of spectral estimates for the random signal displayed in the top panel. The second panel shows a spectrogram with one Hermite taper. Below there is the MT spectrogram with two Hermite windows. In the bottom we show an MT spectrogram with four Hermite windows. The resolution in time and frequency gets worse with the use of more tapers.

serves as a reference. It has a spectral component at $f = 30$ Hz and is centered at $t = 0.3s$. The second signal (middle) is obtained from the reference process by a frequency shift (from $f = 30$ Hz to $f = 100$ Hz). The signal displayed at the bottom is generated by time-shifting the reference process. The spectrogram (middle column) uncovers clearly the time-frequency locations of the signals' spectral contents. The absolute value of the AS in the right column is the same for all three signals.

More generally, one can consider the *filtered Ambiguity spectrum* (Boashash, 2003), which is given by $A_X^{\text{filt}}(\nu,\tau) = A_X(\nu,\tau) \cdot \phi(\nu,\tau)$. The *Ambiguity kernel* $\phi(\nu,\tau)$ is designed to filter out crossterms. Crossterms are an unwanted artefact, which appear when a signal has more than one frequency component. Since they use to appear far from the origin $(\nu,\tau) = (0,0)$, they can be suppressed

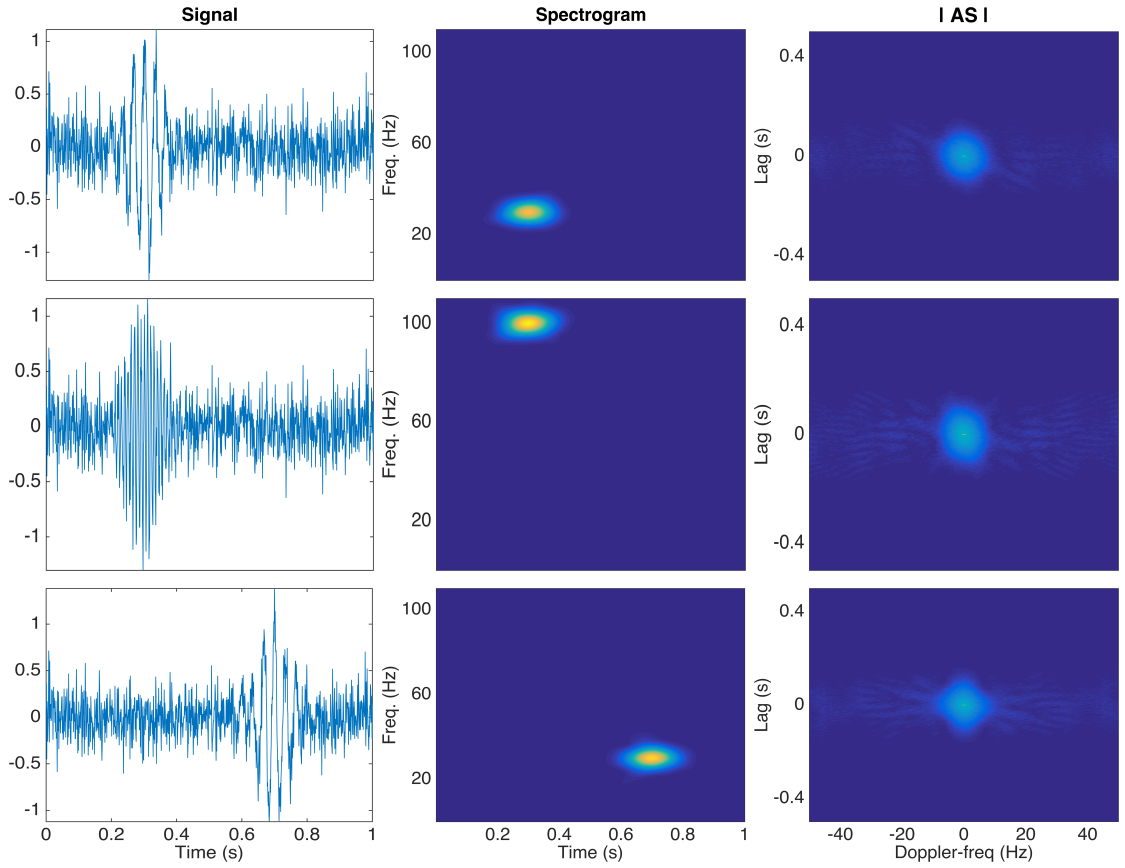


Figure 2.5: The left column displays a reference stochastic process (upper), its frequency shifted version (middle), and its time-shifted version. The middle column shows the spectra estimated by the spectrogram. The right column illustrates the absolute values of the Ambiguity spectra, which are the same for all three signals.

by multiplication with a kernel that is zero for larger values of ν and τ. The effect of filtering is illustrated in Figure 2.6. The stochastic signal displayed in the upper left has frequency components at $t = 0.3s$ and $t = 0.7s$. The lower left panel shows the non-filtered AS, with the desired central term in the origin and the two crossterms on the diagonal. These are eliminated by filtering with the kernel, as illustrated in the lower right panel of Figure 2.6 (with the kernel being displayed in the upper right).

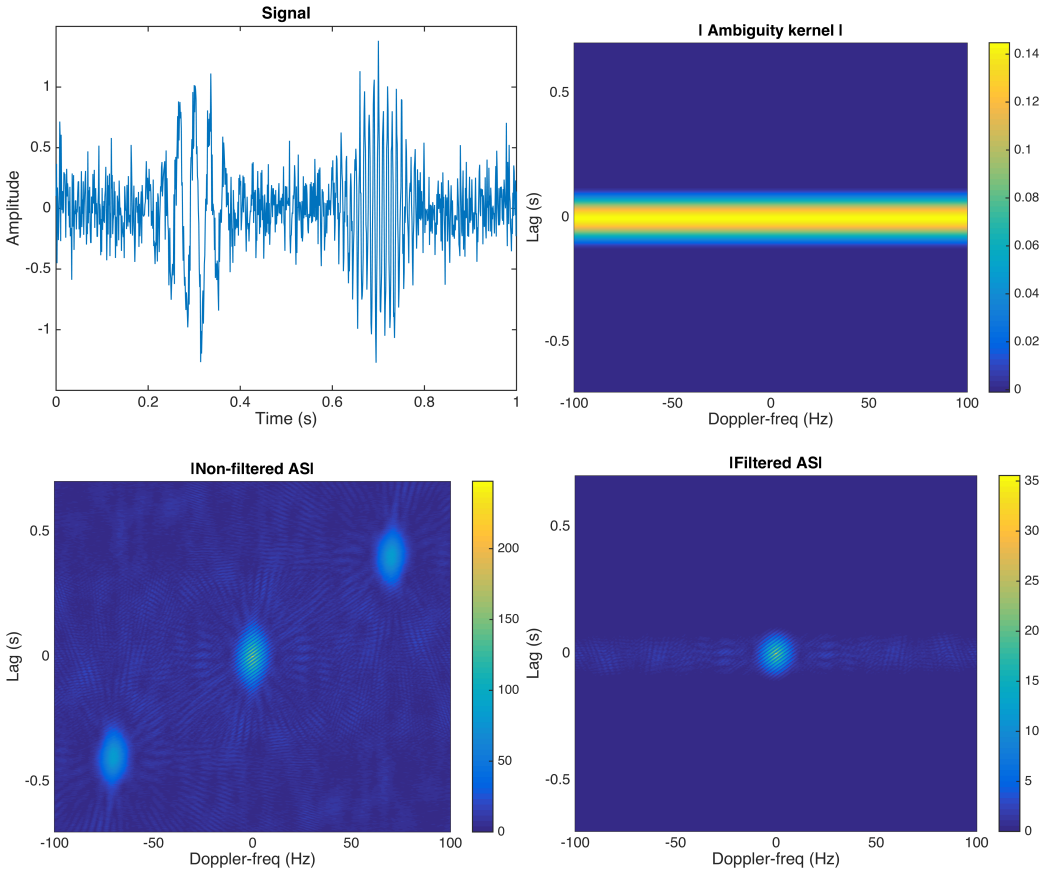


Figure 2.6: Comparison of the (absolute values of the) non-filtered and the filtered AS of a stochastic process shown in the upper left panel. It is the same signal as displayed in Figure 2.1(a), consisting of two Gaussian components. The second component is a time- and frequency-shifted version of the first. The filtered AS in the lower right panel is obtained by multiplication of the non-filtered AS, shown in the lower left, with the Ambiguity kernel as displayed in upper right plot.

Chapter 3

Deterministic and stochastic differential equation models with random effects

This chapter provides background information on likelihood-based inference in deterministic and stochastic differential equation models with random effects for biomedical applications. We start in section 3.1 by reviewing the three sources of noise that are prevalent in biomedical data and discuss how they can be accounted for in the model structure. In section 3.2 we discuss the gains and costs of including different types of random fluctuations in the model, and provide a literature overview on how arising challenges have been addressed. Section 3.3 presents additional information on Paper II, where we motivate the kinetic model, and give some background information on model fitting in Monolix (2016). Section 3.4 embeds the model framework treated in Paper III into a more formal setting and provides a detailed proof on asymptotic properties of the MLE for affine mixed effects, and investigates discretization bias in a simulation study for a non-Lipschitz model.

3.1 Random fluctuations in biomedical data has three main sources

Variability is at the heart of biological systems. It exists on all scales, from the molecular level, over cell processes to organisms and populations, and arises from the complex interplay of numerous system-intern and extern (environmental) factors. Due to the complexity, it is impossible to isolate and explicitly explain all sources of variability. The residual fluctuations in the data, that is, the variability which is not explained by a mechanistic model, are called *noise*. In biomedical applications, measurements are often conducted over time and on several individuals. This design gives rise to three main sources of noise.

Measurement noise, inter-subject variability and intrinsic noise

First of all, there is always some amount of *measurement noise*. This can be caused by improper measurement procedures, changing experimental conditions or observer bias. Measurement noise blurs the signal without changing its dynamics and can be accounted for by appending an additive or multiplicative error term to the model structure.

The second source is the *inter-subject fluctuation*, that is, the unexplained systematic differences of data dynamics between subjects. Individuals share an overall model structure, or base model, but the values of the model parameters differ between subjects. Parts of that inter-subject variability can generally be captured by including subject-specific *covariate information* in the model, such as adjustments for gender, age, body weight or treatment group. However, due to the sheer complexity of real systems, a certain amount of unexplained inter-subject variations will remain. The common way to account for them is by imposing random effects on some (or all) parameters. Models that contain both fixed (parameters that are the same across subjects) and random effects are known as *mixed-effects models* (Sheiner and Beal, 1980; Laird and Ware, 1982; Pinheiro and Bates, 2006). The inference goal is generally the estimation of the fixed parameters (by the distribution mean, mode or median), and the quantification of the extent of random fluctuations (by the variance of the random effects' distribution) using the *whole* data set. The underlying reasoning is that every subject contributes valuable information on the common base model, and by pooling the data one utilizes the entire information in the data to infer the base model (the fixed effects) and to quantify the deviations from it. This approach leads to increased power and robustness of statistical procedures. The challenge in mixed-effects models is that explicit likelihood-based inference is not feasible, because the data likelihood is not explicitly available and has to be approximated.

Lastly, there is the *intrinsic noise*, the unexplained variability within the system itself, such as fluctuations in blood pressure, metabolic processes, or varying stress levels. This type of noise can be substantial in biomedical data, because the underlying data generating process is often too complex to be modeled exactly or is not understood well enough. Such internal random fluctuations can be accounted for by including stochasticity in the dynamical model itself. The dynamics of biological data can often be described by systems of differential equations, and for this the mathematical toolbox offers two kinds of models, (i) the deterministic ordinary differential equations (ODEs) and (ii) their stochastic extension, the stochastic differential equations (SDEs). The driving noise component in SDEs explicitly represents intrinsic system noise and accounts for prevalent model uncertainty or misspecification. As for mixed-effects models, exact maximum

likelihood estimation of parameters is not possible in SDE models (with few exceptions), because the state likelihood is a product of transition densities which are generally unknown. Therefore, also here approximation methods, analytical or numerical, are asked for.

Example 3.1 (The three sources of noise in the data from Paper II)**.**
We exemplify the three types of noise by means of the data set described in Paper II (see also section 3.3), which also is illustrated in Figure 3.1 on page 28. The first to be noticed is that the trajectories show a fairly erratic behavior. This may reflect both *measurement noise*, e.g., caused by movements of the subjects while images were taken, and *system noise*, such as unexplained fluctuations in metabolic processes or breathing rhythm. Another striking observation is the common pattern, i.e., the common base dynamics, that all subjects seem to share. For instance, the counts for the liver compartment have a similar pattern across all participants, with an initial monotone increase up to about $30min$ after administration followed by a monotone, somewhat flatter decrease. However, the trajectories still differ quite substantially across subjects. This *inter-subject variability* may be caused (i) externally, by differences in the subjects' gamma ray attenuations which have not been taken into account properly, or improper identification of the compartments, or (ii) internally, by differences in the subjects' metabolisms, due to factors such as diet, age, weight, or genetic differences.

3.2 Maximum Likelihood inference in ODE and SDE models with random effects

The Maximum Likelihood Estimator (MLE) has a number of desirable properties, such as consistency, asymptotic normality and efficiency. However, as appealing the properties of the MLE may be, it comes with the challenge of an often intractable likelihood.

Balancing model complexity and tractability

When modeling biomedical data, one may easily be tempted to write down a complex system of differential equations, incorporating subject-specific covariate information and, just to be on the safe side, including all possible noise sources, which explain everything that is left unexplained. One may then be carried away by how well the model provides for all contingencies of the real world. While looking good on paper, the model's complexity leads to likelihood intractability and parameter estimation has to be based on (sometimes several) approximations. The sheer appeal of

a complex model which captures "everything", however, bears the pitfall of an overoptimistic faith in the results and subsequent false inference. A reasonable trade-off has to be found between a parsimonious model with reliable and efficient estimation procedures and a complex model, which may capture the data dynamics better, but relies on crude approximations or heavy computations for parameter inference.

When a data set contains a substantial amount of unexplained variability of various types, it undoubtedly is essential to explicitly include these different noise sources. For instance, prevalent intrinsic noise in systems with non-linear dynamics can change the qualitative behavior of the deterministic patterns significantly. In such cases, modeling the dynamics with a deterministic ODE can lead to biased estimates and false inference (Donnet et al., 2010; Bachar et al., 2012; Leander et al., 2014). However, probably more often than not, one source of noise can be assumed to be negligible as compared to others. This observation justifies a simpler model, which dispenses with crude approximations or high computational costs while still allowing for reliable inference.

Illustration: The likelihood in SDMEMs is intractable

For illustration, assume the data generating process X^i for subject i is modeled by a system of differential equations and all three sources of noise are included in the model. We take measurements for subject i at time points $t_k = \frac{k}{n}$ for $k = 1, \dots, n$, and collect all observations in the vector $\boldsymbol{y}$. More specifically, we consider the model

$$\begin{aligned} dX_t^i &= F^i(t, X_t^i, \mu, \phi^i)dt + \Sigma(t, X_t^i)dW_t^i, \quad 0 \leq t \leq 1, \quad X_0^i = x_0^i, \\ y_k^i &= g^i(X_{t_k}^i, \varepsilon_k^i), \quad k = 1, \dots, n. \end{aligned}$$

The dynamics of the underlying signal X^i is thus described by an SDMEM with fixed effect μ, random effect ϕ^i and a Wiener process $W^i = (W_t^i)_{t \in \mathbb{R}}$ to model the intrinsic (subject-specific) noise (see section 3.4 for a more formal description of an SDMEM). At time t_k we observe y_k^i, which is a function of the underlying signal $X_{t_k}^i$ and measurement noise ε_k^i (if $g(x, \varepsilon) = x$, we observe the state directly). The distributions of ϕ^i and ε_k^i are parameterized by ϑ and ξ, respectively, and we assume independence of all noise variables. The target of statistical inference is the estimation of $\theta = (\mu, \vartheta, \xi)$. While the postulated model may in fact account for many contingencies of reality, inference for θ is challenging. The reason is the two-layer intractability of the likelihood

$$p(\boldsymbol{y}|\theta) = \prod_{i=1}^{N} \int \left[\prod_{k=1}^{n} \int p^i(y_k^i | X_{t_k}^i, \phi^i, \theta) p^i(X_{t_k}^i | X_{t_{k-1}}^i, \phi^i, \theta) dX_{t_k}^i \right] p(\phi^i|\theta) d\phi^i.$$

In its inner layer, we find the transition density of the state $p^i(X^i_{t_k}|X^i_{t_{k-1}}, \phi^i, \theta)$, which is unknown for most SDE models. But even if it was known, the integrals over state and random effects in the outer layer are often high-dimensional and lack a closed-form solution.

How likelihood intractability in SDMEMs has been addressed in literature

The unavailability of the transition density in SDE models has spurred much research over the last decades, and neat overviews are Sørensen (2004) and Phillips and Yu (2009). Poulsen (1999) obtains approximate transition densities by employing the result that transition densities are given as solutions to the Kolmogorov forward equations, which are solved numerically. If time steps between observations are small, the transition density is approximately Gaussian. This idea underlies the approaches of Pedersen (1995), Elerian et al. (2001), Eraker (2001), Durham and Gallant (2002) and Golightly and Wilkinson (2008), who use an Euler-Maruyama approximation and combine it with simulation-based data augmentation techniques to ensure closely spaced observation times. Even though this approximation can, in theory, be made arbitrarily exact, it has a fixed discrete-time bias. This drawback is overcome by the so-called exact simulation-based methods considered by Beskos et al. (2006) and Sermaidis et al. (2013). A simulation-free approach is proposed by Aït-Sahalia (2002), who approximates the transition density by a closed-form Hermite expansion. If high-frequency observations are available, one can avoid the transition density altogether by reverting to the continuous-time likelihood (Phillips and Yu, 2009). Asymptotic properties for the resulting estimator have been studied by Yoshida (1992) and Florens-Zmirou (1989).

Several of the methods above have been applied to the SDMEM context. In models *with observation noise*, the state process X^i is unobserved and has to be estimated. One approach for dealing with latent random variables in maximum-likelihood estimation is the expectation-maximization (EM) algorithm (Dempster et al., 1977). To avoid the marginalization over state and random effects, Donnet and Samson (2008) employ the stochastic approximation EM (SAEM) algorithm (Delyon et al., 1999), a stochastic approximation extension of the EM when the E-step is intractable. An alternative and popular approach for recovering the hidden state in SDMEMs with measurement noise is the extended Kalman filter (EKF) (Tornøe et al., 2005; Overgaard et al., 2005; Mortensen et al., 2007; Klim et al., 2009; Delattre and Lavielle, 2013; Leander et al., 2014, 2015). To avoid marginalization over the random effects, Delattre and Lavielle (2013) couple the EKF with the SAEM. A drawback of the EKF is, however, that no theoretical results on its convergence properties are available.

The inclusion of measurement noise adds an extra layer of latency, and thus complexity, to the

model, because it implies that the state itself is unobserved. However, when the measurement noise is small compared to the system noise, it may be excluded from the model for the sake of improved tractability. Nevertheless, key challenges remain: the marginalization over often high-dimensional random effects and the lack of an analytical expression for the transition density. In an SDMEM *without measurement noise*, Picchini et al. (2010) approximate the transition density with Hermite polynomials (Aït-Sahalia, 2002) and the integral over the random effects by Gaussian quadrature. In a subsequent work, they replaced the Gaussian quadrature approach by Laplace approximation to handle multidimensional random effects more efficiently (Picchini and Ditlevsen, 2011). Random effects in the diffusion coefficient are studied by Delattre et al. (2015). They derive the asymptotic normality of the MLE when random effects have an inverse Gamma distribution, given the sampling interval and the number of subjects grow beyond bounds. Delattre et al. (2013) place their work in the frame of high-frequency observations by investigating parameter estimation based on the continuous-time likelihood. For univariate SDMEMs with linear Gaussian mixed effects, they derive the closed-form expression for the continuous-time likelihood, investigate the asymptotic properties of the MLE and quantify the discrete-time bias. Our work in Paper III was motivated by their approach and extends it to a multivariate state process and the inclusion of covariates.

3.3 Supplementary material for Paper II

In Paper II, we use an ODE-based multi-compartment model with random effects and measurement noise for the statistical analysis of pharmacokinetic data. This section presents additional material for the published work, consisting of a heuristic motivation for the chosen model and a short documentation of the estimation process with the software Monolix (2016).

3.3.1 Modeling the data dynamics

Compartmental models

Pharmacokinetics (PK) is the study of drug absorption, distribution and elimination (i.e., metabolism and excretion). An important model class in PK are compartmental models. They describe the concentration of a drug in different compartments of the human body (plasma, or organs such as liver, heart, kidneys) and how it changes over time. The change is characterized by two factors, (i) the order of elimination (zero order, first order, etc.), i.e., in which manner the drug leaves the compartment, and (ii) the elimination (outflow) rate. Most drug elimination processes follow first-

order kinetics. That is, the amount of drug leaving a compartment is proportional to the amount of drug present in that compartment, at that time. Because elimination rates usually depend on a large number of factors (such as body temperature, blood pressure, genetic disposition), many of which can not be identified or modeled, random effects are standard ingredients in PK models.

The data and dynamical model from Paper II

In the study considered in Paper II, eight participants were asked to swallow a liquid containing radio-labeled ^{75}Se-L-selenomethionine (SeMet), after which its amount in stomach, intestine and liver was traced for $2h$ by dynamic camera imaging. On the obtained images, the regions of interest (stomach, intestine, liver) were manually determined by the physicians. The pixels within each of the three compartments were transformed into counts corresponding to the SeMet levels. Additionally, repeated blood samples were taken to trace SeMet counts (transformed from concentration) in the plasma. Figure 3.1 shows the (normalized) data, with colors encoding different subjects. Let $X_S(t), X_I(t), X_L(t), X_P(t)$ denote the underlying SeMet levels in stomach, intestine, liver and plasma, respectively, given in counts. The mathematical model for their dynamics is (cf. Table 1 in Paper II)

$$\begin{aligned} X_S(t) &= \frac{k_a k_e}{C_l(k_a - k_e)} \left(\mathrm{e}^{-k_e t} - \mathrm{e}^{-k_a t}\right) \\ \frac{d}{dt} X_I(t) &= k_1 X_S(t) - k_3 X_I(t) \\ \frac{d}{dt} X_L(t) &= k_3 X_I(t) - k_4 X_L(t) \\ \frac{d}{dt} X_P(t) &= k_2 X_S(t) + k_4 X_L(t-\tau), \end{aligned} \tag{3.1}$$

with $k_1, \ldots, k_4$ being unknown rate parameters and τ an unknown delay.

Heuristic motivation of the model kinetics

We provide a short motivation for the dynamics in (3.1). The statistical model in the article consists of two additional layers, one observational stage, which accounts for overlapping tissues, and one that represents the measurement error model. These are further described in the article and not revisited here.

We first motivate the equation for X_S. In the study, every subject was asked to swallow a liquid containing 29 μg of SeMet, which corresponds to an unknown amount D of counts (the units of the data). We take a simplistic view and imagine that at the start of the experiment, $t = 0$, the whole liquid is located in the mouth and reaches the stomach directly after swallowing. Let

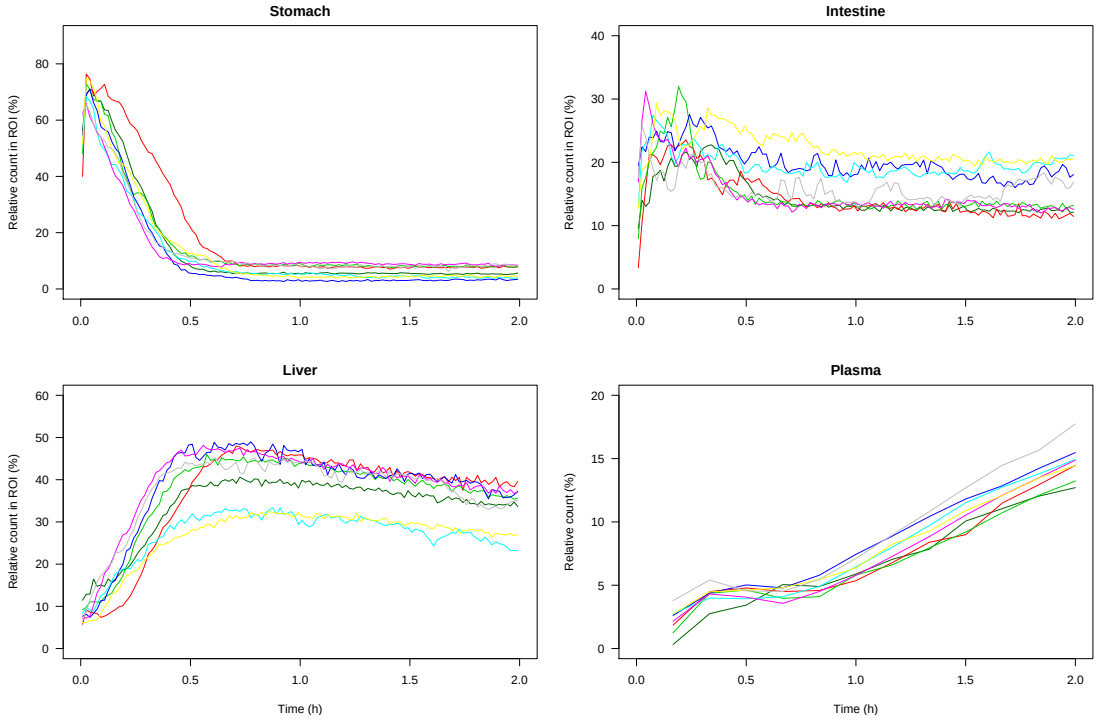


Figure 3.1: The measured data for all participating subjects, with different colors encoding different subjects. The displayed counts were normalized by the total amount of counts present in the body (see Paper II for more information).

$X_M(t)$ be the counts of SeMet in the mouth at time t, and k_a the rate by which they leave the mouth and enter the stomach compartment. Then we have the dynamics $\frac{d}{dt}X_M(t) = -k_a X_M(t)$ with $X_M(0) = D$, or $X_M(t) = D\,\mathrm{e}^{-k_a t}$. We write $X_S(t)$ for the counts of SeMet which are present in the stomach at time t and assume that SeMet leaves the stomach at rate k_e. Then X_S is governed by the differential equation $\frac{d}{dt}X_S(t) = k_a X_M(t) - k_e X_S(t)$, which has the solution $X_S(t) = \frac{k_a \cdot D}{k_a - k_e}\left[\mathrm{e}^{-k_e t} - \mathrm{e}^{-k_a t}\right]$ (assuming $X_S(0) = 0$). The variable C_l in model system (3.1) represents the proportion of dose counts in the stomach that is cleared per hour, $C_l = k_e/D$, such that $X_S(t) = \frac{k_a \cdot k_e}{C_l(k_a - k_e)}\left[\mathrm{e}^{-k_e t} - \mathrm{e}^{-k_a t}\right]$. Note that standard PK models are typically formulated in terms of drug *concentrations*, instead of counts. In that case, D is the total amount of administered dose (29 μg), and C_l is the typical clearance constant, that is, the rate of drug elimination divided by the drug concentration in the stomach compartment, $C_l = k_e/D/V$ (with V being the compartment volume).

The plots in Figure 3.1 reveal that the SeMet level in the stomach reaches its peak immediately after intake. SeMet counts in the intestine compartment do so with a few minutes delay and those

in liver peak thereafter, around $30min$ after administration. This observation is what motivates the cascade-like structure of the proposed model dynamics in (3.1). The plasma compartment is particularly interesting and the delay in (3.1) is crucial for capturing the dynamics in plasma. We observe an initial increase in SeMet levels in plasma during the first $20min$, followed by a $30min$ period of plateauing behavior, after which SeMet levels increase again and steadily. An inflow from the stomach compartment can explain not only the initial increase, but also the plateau, which occurs right about the time when SeMet levels in the stomach have decreased to zero. The post-plateau increase in the plasma data requires therefore an additional source of SeMet inflow, here modeled by a delayed (estimated to ca. $40min$) flow from the liver compartment.
It is noteworthy that this surprisingly simple model provides the best fit among a vast amount of explored models. As pointed out in the article, previous studies have suggested far more complex models. For an illustration, Figure 3.2 shows one version of the models in existing literature. This is an adaptation of the Selenite model from Patterson and Zech (1992, Figure 4), where compartments explicitly related to time courses long after $2h$ (the duration of the study) were omitted. For more information on explored models, how those studies compare to ours, and the conclusions that can be drawn, we kindly refer to our article.

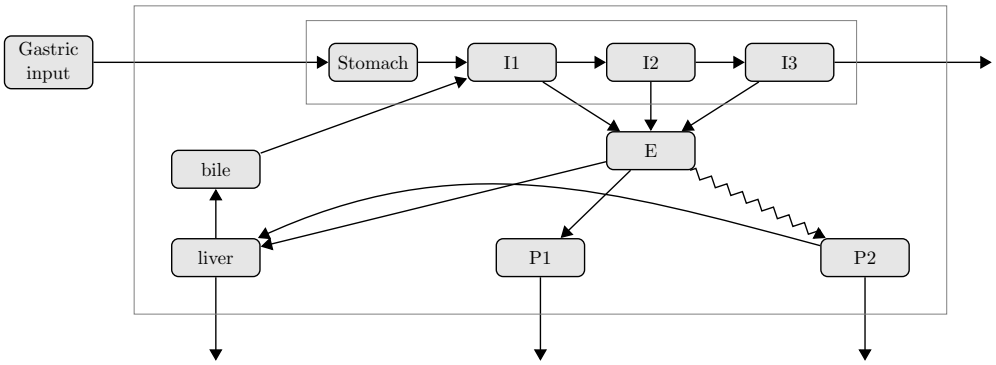


Figure 3.2: Existing literature suggests more complex models. This plot illustrates one of them, adapted to dynamics during the first few hours after drug administration.

3.3.2 Estimation of model parameters with Monolix

For a comprehensive treatment of the practical and theoretical aspects regarding inference with Monolix for PK models, we refer to the book of Lavielle (2014), research articles referenced therein

and to the Monolix website[1]. To assist readers in smoothly getting started with similar models, we give a short description of the key ingredients.

The underlying state dynamics, that is, the hidden and the observational layers of the article model (cf. Table 1 in Paper II) can be specified in Monolix by loading the `.txt`-file shown in Figure 3.3. The measurement error model can easily be chosen in the Monolix GUI. Here one can also indicate which parameters should be equipped with random effects and choose their probability distribution. Because of the substantial (unexplained) variations in the data dynamics between

```
INPUT:
parameter = {a_I, a_L, ka, ke, Cl, k1, k2, k3, k4, tau}

EQUATION:
fL = 1/10          ; fraction of hidden liver out of whole liver
fI = 1/10          ; fraction of hidden intestine out of whole intestine
fPl_S = 1/100      ; fraction of plasma that falls into stomach
fPl_L = 1/10       ; fraction of plasma that falls into liver (hidden & non-hidden)
fPl_I = 1/100      ; fraction of plasma that falls into intestine (hidden & non-hidden)

t0 = 0.00833       ; unit: hours (first measurement taken after 30sec)
XS     = ka*ke / (Cl*(ka-ke))*(exp(-ke*t) - exp(-ka*t))
ddt_XI = k1*XS - k3*XI
ddt_XL = k3*XI - k4*XL
ddt_Pl = k2*XS + k4*delay(XL, tau)

S =  XS + fL*XL + fI* XI + fPl_S*Pl
L = a_L + (1-fL)*(XL + fPl_L*Pl)
I = a_I + (1-fI)*(XI + fPl_I*Pl)

OUTPUT:
output = {L, Pl, I, S}
```

Figure 3.3: Monolix model file that specifies the (hidden layer) of the delay differential equation model equations (hidden stage and observational stage) from Paper II.

participants, parameters of the state dynamics are allowed to vary across subjects and therefore equipped with random effects. We exemplify the random effects modeling and estimation procedure by means of the absorption parameter k_a.

Individual parameters are modeled as transformed Gaussians

Let $\psi^i_{k_a}, i = 1, \ldots, N$, denote the subject-specific absorption parameters, which are independent and identically distributed. In Monolix, it is assumed that a (strictly monotone) transformation of $\psi^i_{k_a}$ is Gaussian distributed, $\phi^i_{k_a} := h(\psi^i_{k_a}) \sim \mathcal{N}(\mu^{\text{pop}}_{k_a}, \sigma^2_{k_a})$. The user specifies the distribution

[1] `http://monolix.lixoft.com/`

of the individual parameters by providing the inverse function h^{-1} to Monolix. In the article, we assumed a log-normal distribution for $\psi^i_{k_a}$ to ensure its positivity. In that case, h^{-1} is the exponential function.

Population parameters are estimated with SAEM

Focusing on k_a, the target of inference is the estimation of the population parameter $\theta_{k_a} = (\mu^{\text{pop}}_{k_a}, \sigma_{k_a})$ by Maximum Likelihood. This is achieved with the SAEM algorithm. Being an iterative algorithm, initial values have to be provided, which is easily done in the GUI. After estimation of θ_{k_a} the MLE for the population rate parameter k^{pop}_a is obtained by $\widehat{k^{\text{pop}}_a} = h^{-1}(\widehat{\mu^{\text{pop}}_{k_a}})$. One can show that $h^{-1}(\mu^{\text{pop}}_{k_a})$ is the median of the distribution of $\psi^i_{k_a}$.

The SAEM version which is implemented in Monolix has two stages (simulated annealing), in order to increase the chances of converging to a global maximum. The fast stage, comprising K_1 iterations, is characterized by erratic estimate trajectories and is used to converge quickly into the area of interest. In the second stage with K_2 iterations, the convergence behavior is much more deterministic. The traces of parameter estimates for the SAEM are plotted and saved in a file `saem.png` (the algorithm convergence graph), and the vector of population estimates together with their covariance structure (given by the estimated Fisher information) are saved in a file named `pop_parameters.txt`.

Some caution is recommended for the interpretation of the standard deviations of the ϕ^i. Since this is the standard deviation on the transformed scale, it is not immediate how it translates to a variation measure on the scale of $\psi^i_{k_a}$. This, however, is important for quantifying how strongly parameters vary across subjects. For log-normal distributions, as in the case of $\psi^i_{k_a}$, the coefficient of variation satisfies $CV = \sqrt{\mathrm{e}^{\sigma^2_{k_a}} - 1}$. If $\sigma^2_{k_a}$ is small, one has $\sigma_{k_a} \approx CV$, such that σ_{k_a} can indeed be used to give a reasonable sense of variation.

The estimated standard deviations of all random effects from the article model are shown in Table 3.1 (excluded from the actual article). In an initial fit, the standard deviation corresponding to $\psi^i_{k_4}$ was not significantly different from 0. Therefore, the model was re-fitted with k_4 as fixed effect. The standard deviations corresponding to the other parameters are significantly larger than zero, justifying the inclusion of random effects for them.

	σ_{a_I}	σ_{a_L}	σ_{k_a}	σ_{k_e}	σ_{C_l}	σ_{k_1}	σ_{k_2}	σ_{k_3}	σ_{k_4}	σ_τ
estimate	0.184	1.582	0.814	0.138	0.286	0.307	0.199	0.490	0	0.138
sd	0.046	0.481	0.220	0.036	0.072	0.079	0.056	0.125	0	0.038

Table 3.1: Estimated standard deviations (and their standard deviations) of the transformed individual parameters $\phi^i = h(\psi^i)$.

Individual parameters are estimated with the Metropolis-Hastings algorithm

After an estimate of θ_{k_a} has been obtained, the individual parameters are estimated with the Metropolis-Hastings (MH) algorithm. For every subject, the MH generates a Markov chain $\{\psi_{k_a}^{i,(m)}, m = 1, \ldots, M\}$, whose stationary distribution is the conditional distribution $p(\psi_{k_a}^i|\boldsymbol{y}^i; \hat{\theta}_{k_a})$. The individual parameter is then estimated by the conditional mean $\widehat{\psi}_{k_a}^{i,\text{mean}} = \frac{1}{M}\sum_{m=1}^{M} \psi_{k_a}^{i,(m)}$ or the conditional mode $\widehat{\psi}_{k_a}^{i,\text{mode}} = \arg\max_{\psi_{k_a}^i} p(\psi_{k_a}^i|\boldsymbol{y}^i, \hat{\theta}_{k_a})$ of that distribution. By default, Monolix uses the conditional mode.

Figures 3.4 - 3.7 display the observed data (solid green) for every subject and compartment. The dotted black lines represent the population fits, i.e., the values obtained when setting the random effects ϕ^i to zero. The individual fits are obtained by using the estimated individual parameters $\widehat{\psi}^i$ in the model and are plotted as dashed red lines. It is evident that by allowing parameters to vary across individuals, the subject-specific dynamics are much better captured.

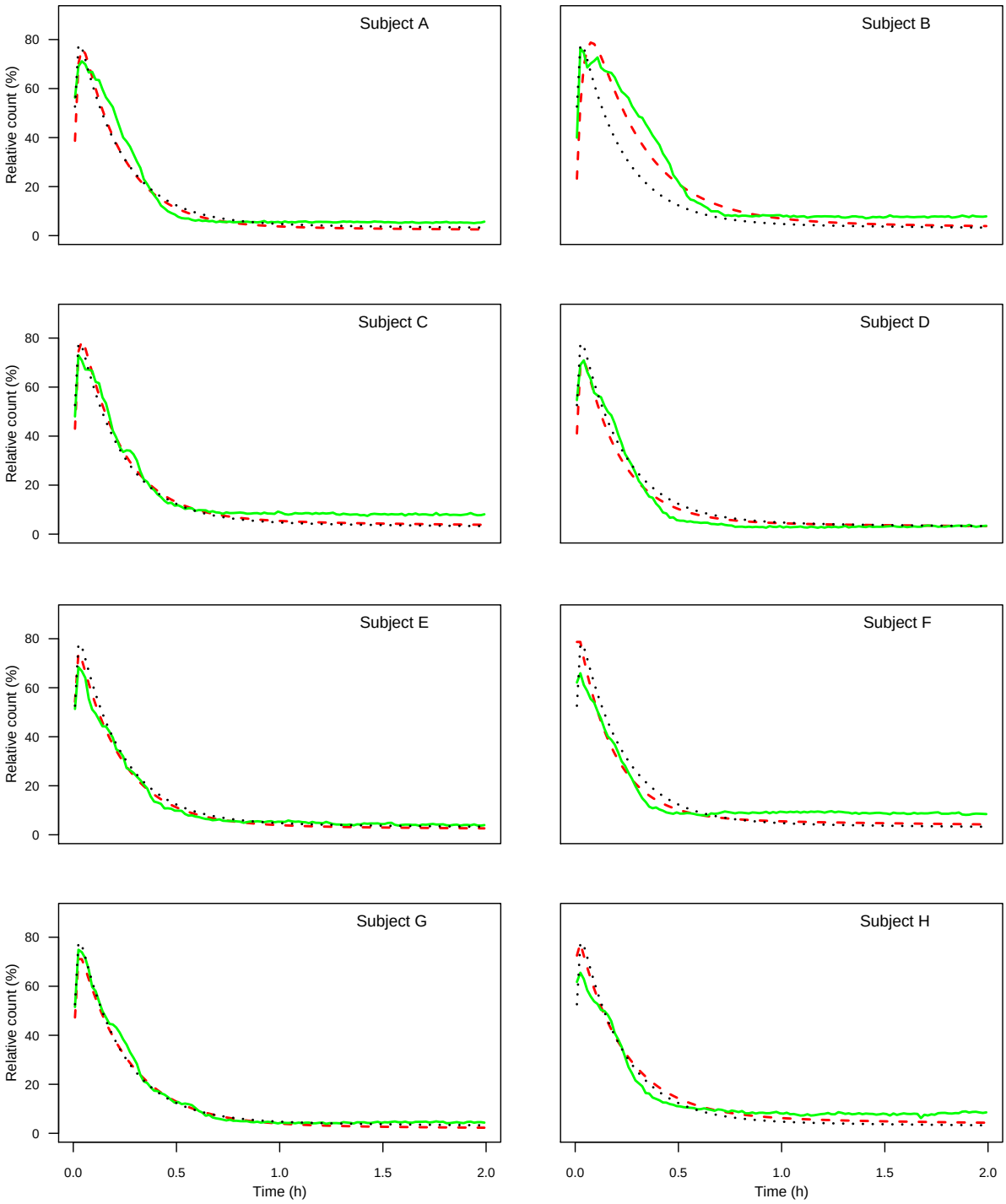


Figure 3.4: Stomach: Observed data (solid green), individual fits (dashed red) and the population fit (dotted black).

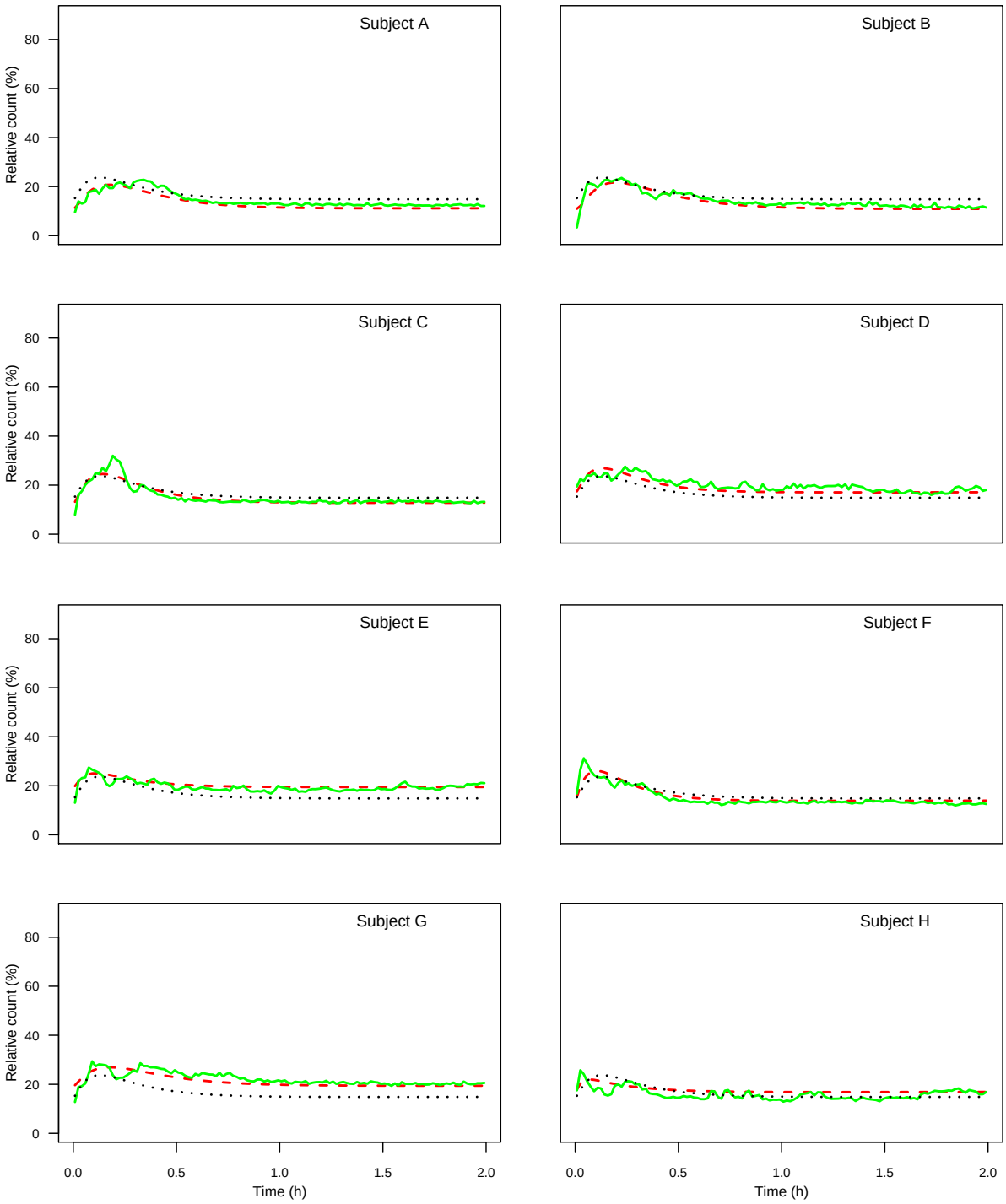


Figure 3.5: Intestine: Observed data (solid green), individual fits (dashed red) and the population fit (dotted black).

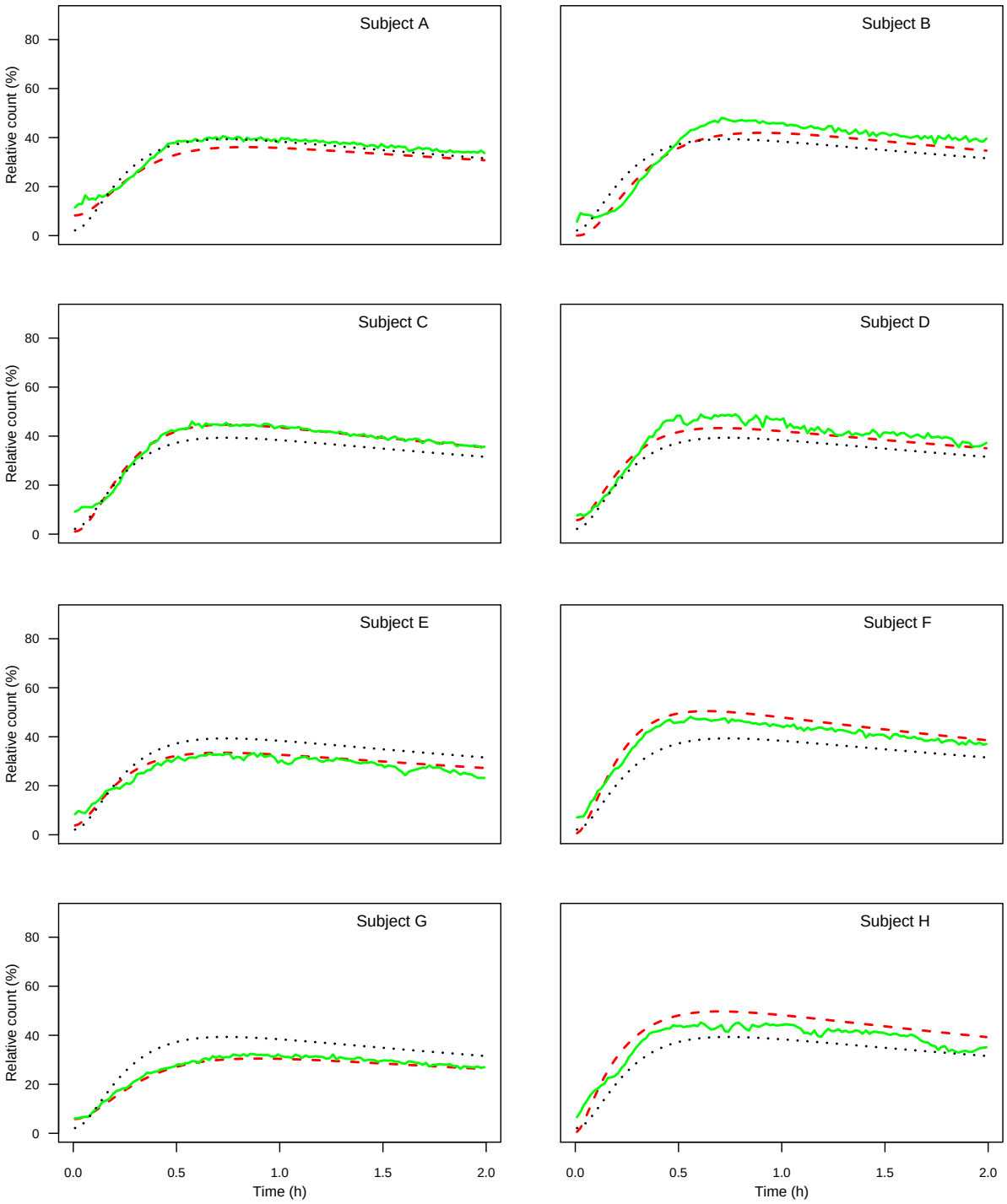


Figure 3.6: Liver: Observed data (solid green), individual fits (dashed red) and the population fit (dotted black).

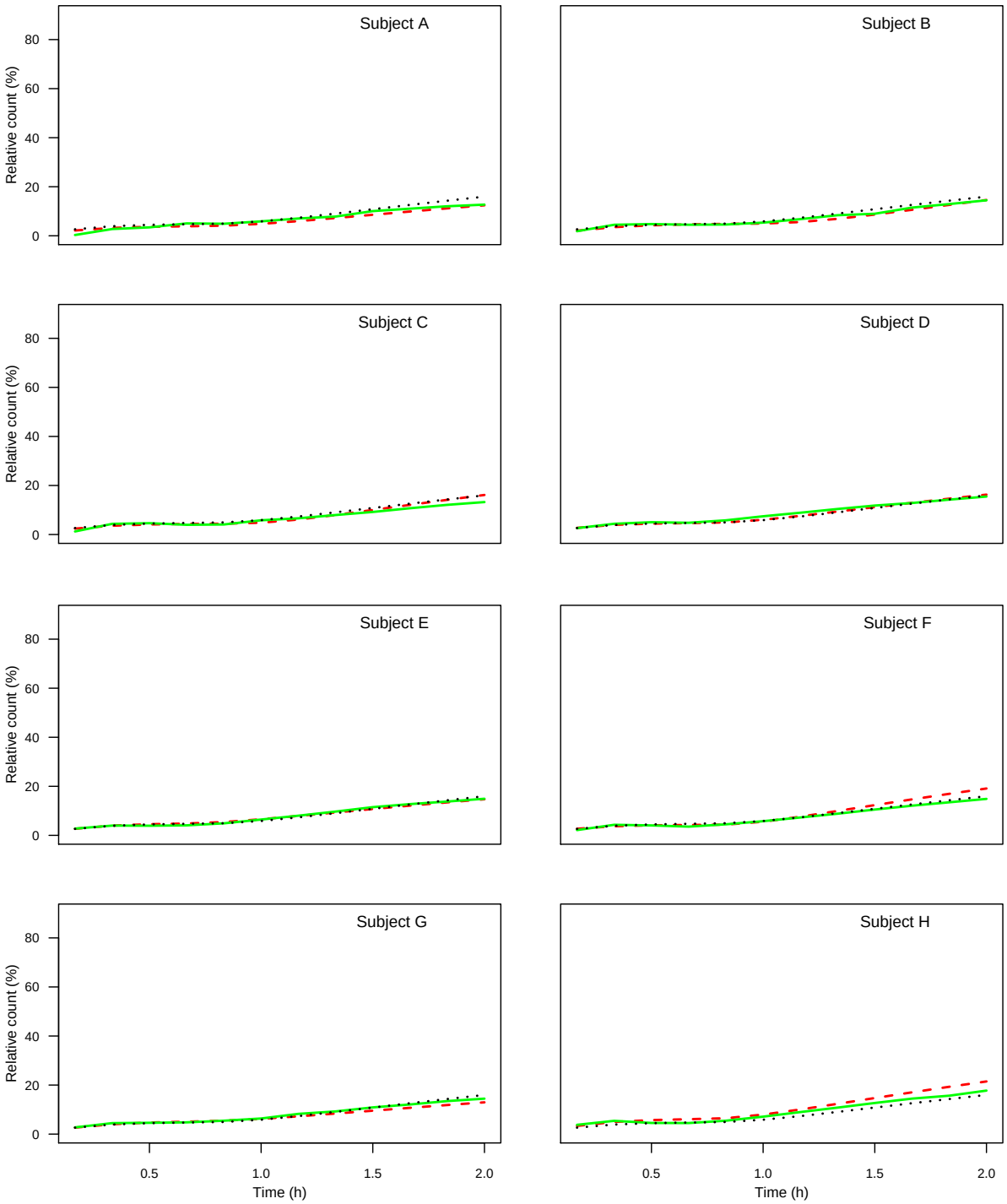


Figure 3.7: Plasma: Observed data (solid green), individual fits (dashed red) and the population fit (dotted black).

3.4 Supplementary material for Paper III

This section embeds the SDMEM framework considered in Paper III into a rigid theoretical setting. We give details on regularity assumptions for the model and provide detailed proofs for asymptotic properties of the MLE when observations are identically distributed and the model is affine in the mixed effects (cf. Section 2.2 in Paper III). This is an extension of the setting considered in Delattre et al. (2013). The proofs, which were excluded from the paper because of fairly cumbersome, but still standard, derivations, follow their ideas. To enable a more convenient reading, we restate the theorems from the article and apologize for some resulting redundancy. The last part of this section is devoted to auxiliary theorems. Some results stated there are adjustments of established theorems to the particular setting considered here and undermined by proofs, while others are well-known results that are presented only for convenience of the reader and left without proof.

Notation We use the symbol $\|\cdot\|$ for the Euclidean vector norm. The $d \times d$ unity matrix is written as I_d. For a $(r \times m)$ matrix A we write A' for its transpose. If A is a square matrix, we write $\mathrm{tr}(A)$ for its trace and $\det(A)$ for its determinant, and if A is invertible, A^{-1} denotes the inverse of A. We will use $[\![\cdot]\!]$ to denote a matrix norm that is (i) sub-multiplicative, i.e., for any two square matrices A_1, A_2 of same dimension one has $[\![A_1 A_2]\!] \leq [\![A_1]\!][\![A_2]\!]$, and (ii) compatible with the Euclidean norm, i.e., $\|Ax\| \leq [\![A]\!]\,\|x\|$. An example is the spectral norm, which is the matrix norm induced by the Euclidean norm, and is defined as $[\![A]\!]_S := \sqrt{\max\{\lambda_1, \ldots, \lambda_m;\}}$, where the λ_i in this definition are the eigenvalues of $A'A$.

3.4.1 Formal definition of SDMEMs

Let $\Theta \subset \mathbb{R}^q$ be a given parameter set, which we assume to be convex and compact. We consider estimation of the *population parameter* $\theta = (\mu, \vartheta) \in \Theta$ for r-dimensional SDMEMs based on data from N independent subjects. In this kind of models it is assumed that the data-generating process for individual i is the solution to an SDE with unknown p-dimensional *fixed effect* μ and unobserved d-dimensional *random effect* ϕ^i,

$$dX_t^i = F^i(t, X_t^i, \mu, \phi^i)dt + \Sigma(t, X_t^i)dW_t^i, \quad 0 \leq t \leq T^i, \quad X_0^i = x_0^i. \tag{3.2}$$

More formally, we assume that $(\Omega, \mathcal{F})$ is a measure space, on which a family of probability measures $\mathcal{P} = \{\mathbb{P}_\theta; \theta \in \Theta\}$ is defined, which is dominated by the probability measure $\mathbb{P}$. The independent r-dimensional Wiener processes $W^i = (W_t^i)_{t \geq 0}$ are defined on $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathcal{P})$, with $\mathcal{F}_t$ being the $\mathbb{P}$-completion of the σ-algebra generated by the $(W_s^i)_{0 \leq s \leq t}$ and ϕ^i, $i = 1, \ldots, N$. The initial conditions

$x_0^i \in \mathbb{R}^r$ and the positive time horizons T^i are known, and we set $T = \max\{T^i, i = 1, \ldots, N\}$. The random effects ϕ^i are independent and identically distributed (i.i.d.) according to a Lebesgue density $g(\cdot; \vartheta)$, with unknown $\vartheta \in \mathbb{R}^{q-p}$, and are independent of the driving Wiener processes. The continuous drift functions $F^i : [0, T] \times \mathbb{R}^{r+p+d} \to \mathbb{R}^r$ and the diffusion coefficient $\Sigma : [0, T] \times \mathbb{R}^r \to \mathbb{R}^r$ are known and measurable, and $\Sigma(t, x)$ is invertible for all $0 \leq t \leq T, x \in \mathbb{R}^r$. By allowing the F^i to vary across individuals, we include the case of not necessarily identically distributed X^i.
The target of statistical inference is Maximum Likelihood estimation of θ from observations of X^i at time points $0 \leq t_0^i < t_1^i < \ldots < t_{n_i}^i \leq T^i$, $i = 1, \ldots, N$. We address the inference task from a continuous-time angle, through studying the MLE derived from the continuous-time likelihood. This is motivated by the high-frequency nature of many data sets that arise in biomedical applications, cf. section 1.2.

Remark 3.2. A useful interpretation of the subject-specific drift functions in the SDMEM framework can be obtained when they are written in the form $F^i(t, X_t^i, \mu, \phi^i) = F(D_t^i, X_t^i, \mu, \phi^i)$, with measurable, deterministic functions $D^i : [0, T] \to \mathbb{R}^s$ and $F : \mathbb{R}^{s+r+d+p} \to \mathbb{R}^r$. The function D^i can then be interpreted as a *covariate* for subject i.

3.4.2 Regularity assumptions for SDMEMs

We need to impose a few regularity assumptions to ensure that the inference problem is well-defined. For better readability, we omit the index i, but it is assumed that the subsequent assumptions hold for all subjects. Under a *solution* to the SDE

$$dX_t = F(t, X_t, \mu, \phi)dt + \Sigma(t, X_t)dW_t, \quad 0 \leq t \leq T, \quad X_0 = x_0, \tag{3.3}$$

we understand an $(\mathcal{F}_t)_t$-adapted, continuous process $X = (X_t)_{t \geq 0}$, which is of the form

$$X_t = x_0 + \int_0^t F(s, X_s, \mu, \phi)ds + \int_0^t \Sigma(s, X_s)dW_s,$$

and has for all $0 \leq t \leq T$ the property $\mathbb{P}\left(\int_0^t \|F(s, X_s, \mu, \phi)\| + [\![\Sigma(s, X_s)]\!]^2 ds < \infty\right) = 1$. Occasionally, we will also consider an SDE where we assume that we actually have observed the value of the random effect, which we may call φ. That is, we will consider a solution to

$$dX_t^\varphi = F(t, X_t^\varphi, \mu, \varphi)dt + \Sigma(t, X_t^\varphi)dW_t, \quad 0 \leq t \leq T, \quad X_0^\varphi = x_0. \tag{3.4}$$

We now state two kinds of assumptions, both of which guarantee the existence of a unique solution to (3.3) (and to (3.4)), which is so well-behaved that certain Radon-Nikodym densities exist. In the

following, generic vectors in $\mathbb{R}^r, \mathbb{R}^p, \mathbb{R}^d$ are denoted by x, μ and φ, respectively, and we let μ_0, φ_0 be fixed.

(A) (i) ϕ has finite moments of any order and there is a positive constant K (that may depend on μ) such that

$$\begin{aligned} \|F(t,x,\mu,\varphi)\| + [\![\Sigma(t,x)]\!] &\leq K\left(1+\|x\|+\|\varphi\|\right) \quad \forall t,x,\varphi \\ \|F(t,x,\mu,\varphi) - F(t,y,\mu,\varphi)\| + [\![\Sigma(t,x)-\Sigma(t,y)]\!] &\leq K\|x-y\| \quad \forall t,x,y,\varphi. \end{aligned}$$

(ii) Let $\Gamma = \Sigma\Sigma'$. Any solution X^φ to (3.4) satisfies
$\mathbb{P}\left(\int_0^T F(t,X_t^\varphi,\mu_0,\varphi_0)'\Gamma(t,X_t^\varphi)^{-1}F(t,X_t^\varphi,\mu_0,\varphi_0)dt < \infty\right) = 1.$

(iii) There are constants $K_\mu, K_\varphi, \kappa_\mu, \kappa_\varphi > 0$, such that

$$\begin{aligned} \left\|[F(t,x,\mu,\varphi) - F(t,x,\mu,\tilde{\varphi})]'\Gamma(t,x)^{-1}\right\| &\leq K_\mu\left(1+\|x\|^{\kappa_\mu}\right)\|\varphi-\tilde{\varphi}\| \quad \forall t,x,\varphi,\tilde{\varphi}, \\ \left\|[F(t,x,\mu,\varphi) - F(t,x,\tilde{\mu},\varphi)]'\Gamma(t,x)^{-1}\right\| &\leq K_\varphi\left(1+\|x\|^{\kappa_\varphi}\right)\|\mu-\tilde{\mu}\| \quad \forall t,x,\mu,\tilde{\mu}. \end{aligned}$$

(B) (i) For all $\theta \in \Theta$, $\|\phi\|$ has a finite moment generating function with respect to $\mathbb{P}_\theta$. The drift function is of the form $F(t,x,\mu,\varphi) = G(t,x,\mu) + C(t,x)\cdot\varphi$, with $G : [0,T]\times\mathbb{R}^{r+p} \to \mathbb{R}^r$, $C : [0,T]\times\mathbb{R}^r \to \mathbb{R}^{r\times d}$. The functions G, C and Σ are uniformly in t Lipschitz-continuous and of sublinear growth in x.

(ii) Assumption (A)(ii) holds.

(iii) There are constants $K, \kappa > 0$, such that for all $t, x, \mu, \tilde{\mu}$,

$$\left\|[G(t,x,\mu) - G(t,x,\tilde{\mu})]'\Gamma(t,x)^{-1}\right\| \leq K(1+\|x\|^\kappa)\|\mu-\tilde{\mu}\|.$$

Assumption (A)(i) is a standard condition for the existence of a unique solution to (3.3). Condition (A)(ii) ensures the existence of a particular Radon-Nikodym density (cf. Theorem 3.4), and (A)(iii) guarantees that this density is product-measurable. This last condition is only needed to ensure continuity of the stochastic integral term in the conditional likelihood in the proof of Theorem 3.5. Note that, when the random effect enters in a multiplicative way as in (B), a sublinear growth condition (cf. (A)(i)) is not necessarily satisfied. A drift function as in (B) is therefore not a special case of (A). However, existence and uniqueness in this case follow from standard SDE theory (cf. Theorem 3.3).

3.4.3 Continuous-time likelihood for general SDMEMs

Theorem 3.3 (Existence and uniqueness of an integrable solution)**.**
Assume that (A)(i) or (B)(i) is satisfied. For every $\theta \in \Theta$ there exists a unique $\mathbb{P}_\theta$-a.s. continuous solution X to (3.3)*. As a consequence, there is also a unique solution X^φ to* (3.4)*. Moreover, X and X^φ satisfy $\sup_{0\le t\le T} \mathbb{E}_{\mathbb{P}_\theta}\left(\|X_t\|^{2k}\right) < \infty$ and $\sup_{0\le t\le T} \mathbb{E}_{\mathbb{P}_\theta}\left(\|X_t^\varphi\|^{2k}\right) < \infty$ for all $k \in \mathbb{N}$.*

Proof of Theorem 3.3. Under (A)(i), the statement is standard and not further considered here. Under (B)(i), the existence of a unique solution follows from Theorem A.1 with $\alpha(t,x,\omega) := F(t,x,\mu,\phi(\omega))$. It only remains to verify the existence of moments. The Lipschitz-continuity and sublinear growth of the functions G and C (cf. condition (B)(i)) assure that there is a constant $K > 0$ (which may depend on μ), such that F satisfies

$$\|F(t,x,\mu,\varphi) - F(t,z,\mu,\varphi)\| \le K\,(1+\|\varphi\|)\,\|x-z\|\,,$$
$$\|F(t,x,\mu,\varphi)\| \le K\,(1+\|\varphi\|)\,(1+\|x\|),$$

i.e., the standard (deterministic) Lipschitz and linear growth conditions with Lipschitz constant $\tilde{L}(\varphi) = K\,(1+\|\varphi\|)$. Let $L(\varphi) = \max\{1, \tilde{L}(\varphi)\}$ and denote by X^φ the unique solution to (3.4). Theorem A.2 gives for any $m = 2k$ the bound

$$h_m(\varphi) := \mathbb{E}_{\mathbb{P}_\theta}\left(\|X_t^\varphi\|^m\right) \le 2^{\frac{m-1}{2}}\,(1+\|x_0\|^m)\,\mathrm{e}^{m(\sqrt{\tilde{L}(\varphi)}+\tilde{L}(\varphi)\frac{m-1}{2})t} \le D_m\,\mathrm{e}^{(m+1)^2 L(\varphi)t}.$$

Let M_θ be the moment-generating function of $\|\phi\|$ under $\mathbb{P}_\theta$. We deduce with Corollary 3.23 in Karatzas and Shreve (1991) (which implies that $X|\phi=\varphi$ has the same distribution as X^φ), that $\mathbb{E}_{\mathbb{P}_\theta}\left(\|X_t\|^m\right) = \mathbb{E}_{\mathbb{P}_\theta}\left(h_m(\phi)\right) \le D_m M_\theta\left((m+1)^2 K t\right) < \infty$. As log-convex function, M_θ is continuous, such that we even have $\sup_{0\le t\le T}\mathbb{E}_{\mathbb{P}_\theta}\left(\|X_t\|^m\right) < \infty$.

□

Heuristically, if X is the solution to (3.3) and $q(\mu,\varphi;X)$ the *conditional likelihood* of X, given that we have observed $\phi = \varphi$, the *unconditional likelihood* should then be obtainable by marginalizing over the random effect, $p(\theta;X) = \int q(\mu,\varphi;X)\,g(\varphi;\vartheta)d\varphi$. For this to yield a well-defined density p, q should, first of all, exist and secondly, be product-measurable. This will be established in the following.

We denote by $(C_T, \mathcal{C}_T)$ the Borel space of continuous $\mathbb{R}^r$-valued functions defined on $[0,T]$, which is associated with the topology of uniform convergence. We write $\mathbb{Q}_{\mu,\varphi}$ for the measure on $(C_T,\mathcal{C}_T)$ induced by the solution to (3.4), and $\mathbb{Q}_\theta$ for the one induced by the solution to (3.3). Expectations

with respect to these two measures are indicated by $\mathbb{E}_{\mu,\varphi}$ and $\mathbb{E}_\theta$, respectively. We from now on assume that (ϕ, X) is the canonical process on $\mathbb{R}^d \times C_T$.

Theorem 3.4 (Conditional likelihood)**.** *Assume that (A) or (B) holds. Then the family* $\{\mathbb{Q}_{\mu,\varphi};\, \mu \in \mathbb{R}^p, \varphi \in \mathbb{R}^d\}$ *of measures is dominated by* $\mathbb{Q}_{\mu_0,\varphi_0}$ *and the Radon-Nikodym derivative is* $\mathbb{Q}_{\mu_0,\varphi_0}$*-a.s. given by*

$$q(\mu,\varphi;X) = exp\left(\int_0^T [F(s,X_s,\mu,\varphi) - F(s,X_s,\mu_0,\varphi_0)]'\Gamma(s,X_s)^{-1}dX_s \right.$$
$$\left. - \frac{1}{2}\int_0^T [F(s,X_s,\mu,\varphi) - F(s,X_s,\mu_0,\varphi_0)]'\Gamma(s,X_s)^{-1}[F(s,X_s,\mu,\varphi) + F(s,X_s,\mu_0,\varphi_0)]\,ds\right).$$

Theorem 3.5 (Product-measurability of the conditional likelihood)**.**
In the setting of Theorem 3.4 there exists $\mathbb{Q}_{\mu_0,\varphi_0}$*-a.s. a continuous version of* $(\mu,\varphi) \mapsto q(\mu,\varphi;X)$*. In particular,* $q(\mu,\varphi;X)$ *is product-measurable.*

Proof of Theorem 3.5. We provide the proof under the assumption that (A) holds. The function

$$f(t,X_t,\mu,\varphi) := [F(t,X_t,\mu,\varphi) - F(t,X_t,\mu_0,\varphi_0)]'\Gamma(t,X_t)^{-1}[F(t,X_t,\mu,\varphi) + F(t,X_t,\mu_0,\varphi_0)]$$

is $\mathbb{Q}_{\mu_0,\varphi_0}$-a.s. continuous in (μ,φ). We therefore conclude from Theorem A.9 the $\mathbb{Q}_{\mu_0,\varphi_0}$-a.s. continuity of $(\mu,\varphi) \mapsto \int_0^T f(t,X_t,\mu,\varphi)dt$, which corresponds to the Lebesgue integral term in $\log q(\mu,\varphi;X)$. In verifying continuity in (μ,φ) of the stochastic integral term in the conditional likelihood, we restrict ourselves to showing the continuity in φ. We write $\varphi \mapsto \int_0^T F(t,X_t,\mu,\varphi)'\Gamma(t,X_t)^{-1}dX_t = I_T(\varphi) + J_T(\varphi)$, with

$$I_T(\varphi) = \int_0^T F(t,X_t,\mu,\varphi)'\Gamma(t,X_t)^{-1}F(t,X_t,\mu_0,\varphi_0)dt,$$
$$J_T(\varphi) = \int_0^T F(t,X_t,\mu,\varphi)'\Gamma(t,X_t)^{-1}[dX_t - F(t,X_t,\mu_0,\varphi_0)dt],$$

and note that continuity of $\varphi \mapsto I_T(\varphi)$ follows as above from Theorem A.9. To prove continuity of $\varphi \mapsto J_T(\varphi)$, we first observe that under $\mathbb{Q}_{\mu_0,\varphi_0}$, $J_t(\varphi), 0 \le t \le T$, is a continuous local martingale. The same holds for the difference process $J_t(\varphi) - J_t(\tilde{\varphi})$ and its quadratic variation is given by

$$\langle J(\varphi) - J(\tilde{\varphi})\rangle_T = \int_0^T [F(t,X_t,\mu,\varphi) - F(t,X_t,\mu,\tilde{\varphi})]'\Gamma(t,X_t)^{-1}[F(t,X_t,\mu,\varphi) - F(t,X_t,\mu,\tilde{\varphi})]\,dt.$$

The Burkholder-Davis-Gundy inequality gives the existence of a constant K such that for any $m \in \mathbb{N}$

$$\mathbb{E}_{\mu_0,\varphi_0}\left(|J_T(\varphi) - J_T(\tilde{\varphi})|^{2m}\right) \leq K\mathbb{E}_{\mu_0,\varphi_0}\left(\langle J(\varphi) - J(\tilde{\varphi})\rangle_T^m\right).$$

Furthermore, due to the sublinear growth condition on Σ in (A)(i) and the sub-multiplicativity of the matrix norm, assumption (A)(iii) implies that there are $K_\mu, \kappa_\mu > 0$, s.t.

$$\left\|[F(t,x,\mu,\varphi) - F(t,x,\mu,\tilde{\varphi})]'\Gamma(t,x)^{-1}\left[F(t,x,\mu,\varphi) - F(t,x,\mu,\tilde{\varphi})\right]\right\|$$

is bounded from above by $K_\mu\left(1 + \|x\|^{\kappa_\mu}\right)\|\varphi - \tilde{\varphi}\|^2$. Applying this bound to the above gives (we see K in the following as a generic constant and allow it to change from one line to another)

$$\begin{aligned}
&\mathbb{E}_{\mu_0,\varphi_0}\left(|J_T(\varphi) - J_T(\tilde{\varphi})|^{2m}\right)\\
&\quad\leq K\mathbb{E}_{\mu_0,\varphi_0}\left(\langle J(\varphi) - J(\tilde{\varphi})\rangle_T^m\right)\\
&= K\,\mathbb{E}_{\mu_0,\varphi_0}\left(\left[\int_0^T [F(s,X_s,\mu,\varphi) - F(s,X_s,\mu,\tilde{\varphi})]'\Gamma(s,X_s^i)^{-1}\left[F(s,X_s,\mu,\varphi) - F(s,X_s,\mu,\tilde{\varphi})\right]ds\right]^m\right)\\
&\leq K\|\tilde{\varphi} - \varphi\|^{2m}\,\mathbb{E}_{\mu_0,\varphi_0}\left(\int_0^T\left(1 + \|X_s\|^{2m(\kappa_\mu+1)}\right)ds\right)\\
&= K\|\tilde{\varphi} - \varphi\|^{2m}\left(1 + \int_0^T\mathbb{E}_{\mu_0,\varphi_0}\left(\|X_s\|^{2m(\kappa_\mu+1)}\right)ds\right)\\
&\leq K\|\tilde{\varphi} - \varphi\|^{2m}.
\end{aligned}$$

Choosing $2m > d$, the Kolmogorov criterion (Theorem A.5) assures that $\varphi \mapsto J_T(\varphi)$ admits a continuous version $\mathbb{Q}_{\mu_0,\varphi_0}$-a.s. Being continuous in (μ,φ) and measurable in the other argument, the joint measurability of the conditional likelihood follows.

□

The unconditional likelihood is a consequence of Corollary 3.23 in Karatzas and Shreve (1991) and Fubini's theorem.

Theorem 3.6 (Unconditional likelihood)**.**
Assume that (A) or (B) is satisfied. Then $\mathbb{Q}_\theta$ admits a density with respect to $\mathbb{Q}_{\mu_0,\varphi_0}$, which is $\mathbb{Q}_{\mu_0,\varphi_0}$-a.s. given by

$$p(\theta;X) = \frac{d\mathbb{Q}_\theta}{d\mathbb{Q}_{\mu_0,\varphi_0}}(X) = \int_{\mathbb{R}^d} q(\mu,\varphi;X)\,g(\varphi;\vartheta)\,d\varphi. \tag{3.5}$$

Proof. The measures $\mathbb{Q}_\theta, \mathbb{Q}_{\varphi_0}$ are probability measures on the space $(C_T, \mathcal{C}_T)$. With $\mathbb{1}_A$ as the indicator function of a set $A \in \mathcal{C}_T$ one has to show that $\mathbb{E}_\theta(\mathbb{1}_A) = \mathbb{E}_{\mu_0,\varphi_0}(p(\theta;X)\mathbb{1}_A)$. This, however, follows immediately. Because of Corollary 3.23 in Karatzas and Shreve (1991), we have

$$\mathbb{E}_\theta(\mathbb{1}_A) = \int_{\mathbb{R}^d} \mathbb{Q}_{\mu,\varphi}(A) g(\varphi;\theta) d\varphi$$

and by definition of q (cf. Theorem 3.4),

$$= \int_{\mathbb{R}^d} \left[\int_{C_T} \mathbb{1}_A(X) q(\mu,\varphi;X) d\mathbb{Q}_{\mu_0,\varphi_0}(X) \right] g(\varphi;\vartheta) d\varphi.$$

Fubini allows us to change integration order,

$$= \int_{C_T} \mathbb{1}_A(X) \int_{\mathbb{R}^d} q(\mu,\varphi;X) g(\varphi;\vartheta) d\varphi \; d\mathbb{Q}_{\mu_0,\varphi_0}(X),$$

and we obtain the result

$$= \int_{C_T} \mathbb{1}_A(X)\, p(\theta;X)\, d\mathbb{Q}_{\mu_0,\varphi_0}(X) = \mathbb{E}_{\mu_0,\varphi_0}(p(\theta;X)\mathbb{1}_A).$$

□

The extension to the N-sample likelihood is now a direct consequence of Theorem 3.6. We denote by $\mathbb{Q}^i_\theta$ the distribution on $(C_{T^i}, \mathcal{C}_{T^i})$ induced by the solution to (3.2) and by $\mathbb{Q}^i_{\mu,\varphi}$ the one given $\phi^i = \varphi$. We let (ϕ^i, X^i) be the canonical processes on $\mathbb{R}^d \times C_{T^i}$.

Theorem 3.7 (Unconditional N-sample likelihood)**.**
For every $i = 1, \dots, N$, we assume that F^i and Σ are such that (A) or (B) is satisfied. Then the product measure $\mathbb{Q}^{\otimes N}_\theta = \bigotimes_{i=1}^N \mathbb{Q}^i_\theta$ has a density with respect to $\mathbb{Q}^{\otimes N}_{\mu_0,\varphi_0} = \bigotimes_{i=1}^N \mathbb{Q}^i_{\mu_0,\varphi_0}$, which is $\mathbb{Q}^{\otimes N}_{\mu_0,\varphi_0}$-a.s. given by

$$p_N(\theta; X^1, \dots, X^N) = \frac{d\mathbb{Q}^{\otimes N}_\theta}{d\mathbb{Q}^{\otimes N}_{\mu_0,\varphi_0}}(X^1, \dots, X^N) = \prod_{i=1}^N p^i(\theta; X^i) = \prod_{i=1}^N \int_0^{T^i} q^i(\mu,\varphi;X^i) g(\varphi;\vartheta) d\varphi. \tag{3.6}$$

3.4.4 Continuous-time likelihood and MLE for affine Gaussian SDMEMs

Assume the individual drifts are as in (B), that is, the random effects enter the drift in an affine manner. Note that the dependence of the drift on the fixed effect does not need to be affine. Then

the conditional likelihood is $q^i(\mu,\varphi;X^i) = \mathrm{e}^{U_{1i}(\mu)+\varphi' U_{2i}-\frac{1}{2}V_{1i}(\mu)-\frac{1}{2}\varphi' V_{2i}\varphi-\varphi' Z_i(\mu)}$, with

$$\begin{aligned}
U_{1i}(\mu) &= \int_0^{T^i} \left[G^i(t,X_t^i,\mu)-G^i(t,X_t^i,\mu_0)\right]'\Gamma(t,X_t^i)^{-1}\left[dX_t^i - G^i(t,X_t^i,\mu_0)dt\right],\\
V_{1i}(\mu) &= \int_0^{T^i} \left[G^i(t,X_t^i,\mu)-G^i(t,X_t^i,\mu_0)\right]'\Gamma(t,X_t^i)^{-1}\left[G^i(t,X_t^i,\mu)-G^i(t,X_t^i,\mu_0)\right]dt,\\
Z_i(\mu) &= \int_0^{T^i} C^i(t,X_t^i)'\Gamma(t,X_t^i)^{-1}\left[G^i(t,X_t^i,\mu)-G^i(t,X_t^i,\mu_0)\right]dt.\\
U_{2i} &= \int_0^{T^i} C^i(t,X_t^i)'\Gamma(t,X_t^i)^{-1}\left[dX_t^i - G^i(t,X_t^i,\mu_0)dt\right],\\
V_{2i} &= \int_0^{T^i} C^i(t,X_t^i)'\Gamma(t,X_t^i)^{-1}C^i(t,X_t^i)dt.
\end{aligned}$$

If we further assume that the random effects are Gaussian, $g(\varphi;\Omega)=\mathcal{N}(0,\Omega)$, the particular form of the q^i implies that (3.6) turns into an explicit expression

$$p^i(\theta;X^i) = \frac{1}{\sqrt{\det(I_d+V_{2i}\Omega)}}\mathrm{e}^{U_{1i}(\mu)-\frac{1}{2}V_{1i}(\mu)+[Z_i(\mu)+U_{2i}(\mu)]'(V_{2i}+\Omega^{-1})^{-1}[Z_i(\mu)+U_{2i}(\mu)]} \qquad \mathbb{Q}^i_{\mu_0,0}\text{-a.s.}$$

If we even further assume that also the fixed effect enters the drift in an affine manner, the individual log-likelihoods are quadratic in μ and the MLE $\hat{\mu}_N$ becomes explicit.

Theorem 3.8 (Unconditional likelihood for affine Gaussian SDMEMs)**.** *Let the drift functions in* (3.2) *be affine in the fixed and the random effect,* $F^i(t,x,\mu,\phi^i) = A^i(t,x)+B^i(t,x)\mu+C^i(t,x)\phi^i$*, and satisfy (B). Assume additionally that the random effects are Gaussian distributed,* $g(\varphi;\Omega)=\mathcal{N}(0,\Omega)(\varphi)$*. Then the conditional likelihood for subject* i *is* $\mathbb{Q}^i_{0,0}$*-a.s. given by* $q^i\left(\mu,\varphi;X^i\right) = e^{\mu' U_{1i}-\frac{1}{2}\mu' V_{1i}\mu+\varphi' U_{2i}-\frac{1}{2}\varphi' V_{2i}\varphi-\varphi' Z_i\mu}$*, with sufficient statistics*

$$\begin{aligned}
U_{1i} &= \int_0^{T^i} B^i(t,X_t^i)'\Gamma(t,X_t^i)^{-1}\left[dX_t^i - A^i(t,X_t^i)dt\right],\\
V_{1i} &= \int_0^{T^i} B^i(t,X_t^i)'\Gamma(t,X_t^i)^{-1}B^i(t,X_t^i)dt,\\
U_{2i} &= \int_0^{T^i} C^i(t,X_t^i)'\Gamma(t,X_t^i)^{-1}\left[dX_t^i - A(t,X_t^i)dt\right],\\
V_{2i} &= \int_0^{T^i} C^i(t,X_t^i)'\Gamma(t,X_t^i)^{-1}C^i(t,X_t^i)dt,\\
Z_i &= \int_0^{T^i} C^i(t,X_t^i)'\Gamma(t,X_t^i)^{-1}B^i(t,X_t^i)dt.
\end{aligned} \tag{3.7}$$

Integration over φ yields the unconditional likelihood for subject i, which is $\mathbb{Q}^i_{0,0}$-a.s. equal to

$$p^i(\theta; X^i) = \frac{1}{\sqrt{det(I_d + V_{2i}\Omega)}} exp\left(\left[U'_{1i} - U'_{2i}R^i(\Omega)Z_i\right]\mu - \frac{1}{2}\mu'\left[V_{1i} - Z'_iR^i(\Omega)Z_i\right]\mu + \frac{1}{2}U'_{2i}R^i(\Omega)U_{2i}\right), \tag{3.8}$$

with $R^i(\Omega) = (V_{2i} + \Omega^{-1})^{-1}$.

Remark 3.9 (Only fixed effects)**.** If the model contains only fixed effects, one has $p^i(\theta; X^i) = \mathrm{e}^{\mu' U_{1i} - \frac{1}{2}\mu' V_{1i}\mu}$ and the MLE is given by $\hat{\mu}_N = \left[\sum_{i=1}^N V_{1i}\right]^{-1} \sum_{i=1}^N U_{1i}$.

Theorem 3.10 (Maximum Likelihood estimator)**.** *Assume that V_{2i} is strictly positive definite and let $\ell_N(\theta; \boldsymbol{X}) = \log p_N(\theta; \boldsymbol{X})$ denote the log-likelihood and $\boldsymbol{X} = (X^1, \ldots, X^N)$. The score equations are*

$$\begin{aligned}
\frac{d}{d\mu}\ell_N(\theta; \boldsymbol{X}) &= \sum_{i=1}^N \left[U'_{1i} - U'_{2i}R^i(\Omega)Z_i - \mu'\left[V_{1i} - Z'_iR^i(\Omega)Z_i\right]\right] \\
\frac{d}{d\Omega}\ell_N(\theta; \boldsymbol{X}) &= -\frac{1}{2}\sum_{i=1}^N \left[G^i(\Omega) - \gamma^i(\hat{\theta}_N)\gamma^i(\hat{\theta}_N)'\right],
\end{aligned}$$

with $R^i(\Omega) = \left(V_{2i} + \Omega^{-1}\right)^{-1}$ as defined above, $G^i(\Omega) = (I_d + V_{2i}\Omega)^{-1}V_{2i}$, and $\gamma^i(\theta) = G^i(\Omega)V_{2i}^{-1}(U_{2i} - Z_i\mu)$. The MLE $\hat{\theta}_N = (\hat{\mu}_N, \hat{\Omega}_N)$ of the population parameters is given by the system

$$\begin{aligned}
\hat{\mu}_N &= \left[\sum_{i=1}^N \left[V_{1i} - Z'_iR^i(\hat{\Omega}_N)Z_i\right]\right]^{-1} \left[\sum_{i=1}^N \left[U_{1i} - Z'_iR^i(\hat{\Omega}_N)U_{2i}\right]\right] \\
&\sum_{i=1}^N \left[G^i(\hat{\Omega}_N) - \gamma^i(\hat{\theta}_N)\gamma^i(\hat{\theta}_N)'\right] = 0.
\end{aligned} \tag{3.9}$$

Remark 3.11. Note that, as Ω is strictly positive definite, the positive definiteness of V_{2i} implies the positive definiteness of $V_{2i} + \Omega^{-1}$ and positive semi-definiteness of $V_{2i}\Omega$. This ensures that $I_d + V_{2i}\Omega$ is invertible, such that $G^i(\Omega)$ is indeed well-defined. Moreover, $G^i(\Omega)$ is symmetric, which can be seen from the identity $G^i(\Omega) = V_{2i} + V_{2i}(\Omega^{-1} + V_{2i})^{-1}V_{2i}$.

Proof of Theorem 3.10. The log-likelihood is given by

$$\begin{aligned}\ell(\theta; X^i) = \frac{1}{2}\log\left([\det(I_d + V_{2i}\Omega)]^{-1}\right) + U_{1i}'\mu - U_{2i}'R^i(\Omega)Z_i\mu \\ - \frac{1}{2}\mu' V_{1i}\mu + \frac{1}{2}\mu' Z_i' R^i(\Omega) Z_i\mu + \frac{1}{2}U_{2i}' R^i(\Omega) U_{2i}\end{aligned}$$

and we observe that

$$\log\left([\det(I_d + V_{2i}\Omega)]^{-1}\right) = \log\left(\det([I_d + V_{2i}\Omega]^{-1})\right) = \log\left(\det(\Omega^{-1}R^i(\Omega))\right).$$

Differentiation of $\ell(\theta; X^i)$ with respect to μ is straightforward. To compute the derivative with respect to Ω, we recall a few facts from matrix calculus (Petersen and Pedersen, 2008).

(a) The trace is invariant to cyclic permutations.

(b) For two matrices A, Y of suitable dimensions, we have

$$\partial\mathrm{tr}(Y) = \mathrm{tr}(\partial Y), \quad \partial(AY) = A\partial Y, \quad \partial\log\det(Y) = \mathrm{tr}(Y^{-1}\partial Y), \quad \partial Y^{-1} = -Y^{-1}(\partial Y)Y^{-1}.$$

(c) Let f be a scalar-valued function with matrix argument. Then $= f(\Omega) = \mathrm{tr}\left(f(\Omega)\right)$, and its differential and derivative are related via $\partial f(\Omega) = \mathrm{tr}\left(\frac{d}{d\Omega}f(\Omega)(\partial\Omega)\right)$.

We first consider differentiation with respect to Ω of the logarithmic term. With (b) we have

$$\partial\log\left(\det(\Omega^{-1}R^i(\Omega))\right) = \mathrm{tr}\left(R^i(\Omega)^{-1}\Omega\ \ \partial(\Omega^{-1}R^i(\Omega))\ \right).$$

The product rule gives $\partial(\Omega^{-1}R^i(\Omega)) = \Omega^{-1}(\partial R^i(\Omega)) + (\partial\Omega^{-1})R^i(\Omega)$ and (b) implies

$$(\partial\Omega^{-1}) = -\Omega^{-1}(\partial\Omega)\Omega^{-1} \quad \text{and} \quad \partial R^i(\Omega) = R^i(\Omega)\Omega^{-1}(\partial\Omega)\Omega^{-1}R^i(\Omega).$$

Combining these yields

$$\begin{aligned}\partial(\Omega^{-1}R^i(\Omega)) &= \Omega^{-1}(\partial R^i(\Omega)) + (\partial\Omega^{-1})R^i(\Omega) = -\left[\Omega^{-1} - \Omega^{-1}R^i(\Omega)\Omega^{-1}\right](\partial\Omega)\Omega^{-1}R^i(\Omega) \\ &= -G^i(\Omega)(\partial\Omega)\Omega^{-1}R^i(\Omega),\end{aligned}$$

such that we arrive at

$$\partial\log\left(\det(\Omega^{-1}R^i(\Omega))\right) = -\mathrm{tr}\left(R^i(\Omega)^{-1}\Omega\, G^i(\Omega)(\partial\Omega)\Omega^{-1}R^i(\Omega)\right) = -\mathrm{tr}\left(G^i(\Omega)\,\partial\Omega\right),$$

and (c) gives us $\frac{d}{d\Omega}\log\left(\det(\Omega^{-1}R^i(\Omega))\right) = -G^i(\Omega)$. The remaining terms can be treated similarly (using the invariance of the trace to cyclic permutations),

$$\begin{aligned}
\partial\left(U_{2i}'R^i(\Omega)Z_i\mu\right) &= \frac{1}{2}\partial\left(U_{2i}'R^i(\Omega)Z_i\mu\right) + \frac{1}{2}\partial\left(\mu' Z_i' R^i(\Omega)U_{2i}\right) \\
&= \frac{1}{2}\left[\operatorname{tr}\left(U_{2i}'\left(\partial R^i(\Omega)\right)Z_i\mu\right)\operatorname{tr}\left(\mu' Z_i'\left(\partial R^i(\Omega)\right)U_{2i}\right)\right] \\
&= \frac{1}{2}\operatorname{tr}\left(\Omega^{-1}R^i(\Omega)\left[Z_i\mu U_{2i}' + U_{2i}\mu' Z_i'\right]R^i(\Omega)\Omega^{-1}(\partial\Omega)\right) \\
\partial\left(\mu' Z_i' R^i(\Omega)Z_i\mu\right) &= \operatorname{tr}\left(\Omega^{-1}R^i(\Omega)Z_i\mu\mu' Z_i' R^i(\Omega)\Omega^{-1}(\partial\Omega)\right) \\
\partial\left(U_{2i}'R^i(\Omega)U_{2i}\right) &= \operatorname{tr}\left(\Omega^{-1}R^i(\Omega)U_{2i}U_{2i}'R^i(\Omega)\Omega^{-1}(\partial\Omega)\right).
\end{aligned}$$

We apply (c) to these equalities, use $\Omega^{-1}R^i(\Omega) = G^i(\Omega)V_{2i}^{-1}$ and $R^i(\Omega)\Omega^{-1} = V_{2i}^{-1}G^i(\Omega)$, and obtain

$$\frac{d}{d\Omega}\ell(\theta;X^i) = -\frac{1}{2}G^i(\Omega) - \frac{1}{2}G^i(\Omega)V_{2i}^{-1}\left[\left(Z_i\mu U_{2i}' + U_{2i}\mu' Z_i'\right) - Z_i\mu\mu' Z_i' - U_{2i}U_{2i}'\right]V_{2i}^{-1}G_i(\Omega).$$

The bracket term in the middle can be written as $-(U_{2i}-Z_i\mu)(U_{2i}-Z_i\mu)'$, giving

$$= -\frac{1}{2}G^i(\Omega) + \frac{1}{2}G^i(\Omega)V_{2i}^{-1}(U_{2i}-Z_i\mu)(U_{2i}-Z_i\mu)'V_{2i}^{-1}G^i(\Omega).$$

With $\gamma^i(\theta) = G^i(\Omega)V_{2i}^{-1}(U_{2i}-Z_i\mu)$ as introduced above, we arrive at the neat expression

$$\frac{d}{d\Omega}\ell(\theta;X^i) = -\frac{1}{2}\left[G^i(\Omega) - \gamma^i(\theta)\gamma_i(\theta)'\right].$$

□

Discrete data

The high-frequency nature of many biomedical data sets justifies a continuous-time approach. However, the assumption to observe the entire paths $(X_t^i)_{0\le t\le T}$ is still an approximation. Theorem 1 in Paper III states a result on the approximation error, when we observe $X^{i,(n)} := (X_{t_0}^i,\ldots,X_{t_n}^i)$ at time points $t_k = \frac{k}{n}T$ and the integrals in q^i are replaced by finite sums. A Lebesgue integral is approximated by Riemann sums and a stochastic integral of the form $\int_{t_k}^{t_{k+1}} h(s,X_s^i)dX_s^i$ is replaced $h(t_k,X_k^i)\Delta X_k^i$.

Theorem 3.12 (Negligibility of discretization error)**.**
Assume that $A, B'\Gamma^{-1}B, B'\Gamma^{-1}C, C'\Gamma^{-1}C, B'\Gamma^{-1}, C'\Gamma^{-1}$ are globally Lipschitz-continuous in (t,x) and that in addition to A,B,C and Σ also $B'\Gamma^{-1}, C'\Gamma^{-1}$ is of sublinear growth in x, uniformly in t. Then, for all $p\ge 1$ and all $i=1,\ldots,N$, there is a constant K such that

$$\mathbb{E}_\theta\left([\![V_{1i}-V_{1i}^n]\!]^p + \|U_{1i}-U_{1i}^n\|^p + [\![V_{2i}-V_{2i}^n]\!]^p + \|U_{2i}-U_{2i}^n\|^p + \|Z_i-Z_i^n\|^p\right) \le K\left(\frac{T}{n}\right)^{p/2}.$$

Proof of Theorem 3.12. We will restrict ourselves to verifying the bounds on U_{2i}, V_{2i}, $\mathbb{E}_{\theta_0}\left([\![U_{2i} - U_{2i}^n]\!]^p\right) + \mathbb{E}_\theta\left([\![V_{2i} - V_{2i}^n]\!]^p\right) \leq Kn^{-p/2}$. For simplicity, we have included the factor $T^{p/2}$ in the constant K, and in the following will allow K to change from line to line. The bounds for the discretization error in U_{1i}, V_{1i} and Z_i can be derived analogously.

To start with, we verify that $\mathbb{E}_\theta\left(\left\|X_{t+h}^i - X_t^i\right\|^p\right) \leq Kh^{p/2}$ holds for any $0 \leq t \leq t+h \leq T, 0 < h < 1$.

$$\begin{aligned}
\left\|X_{t+h}^i - X_t^i\right\|^p &= \left\| \int_t^{t+h} \left[A(s, X_s^i) + B(s, X_s^i)\mu + C(s, X_s^i) \cdot \phi^i\right] ds + \int_t^{t+h} \Sigma(s, X_s^i) dW_s^i \right\|^p \\
&\leq K \left(\int_t^{t+h} \left\|A(s, X_s^i)\right\| ds \right)^p + K \left(\int_t^{t+h} \left\|B(s, X_s^i) \cdot \mu\right\| ds \right)^p \\
&\quad + K \left(\int_t^{t+h} \left\|C(s, X_s^i) \cdot \phi^i\right\| ds \right)^p + K \left\| \int_t^{t+h} \Sigma(s, X_s^i) dW_s^i \right\|^p .
\end{aligned}$$

We use the general relation $\|Ax\| \leq K\left([\![A]\!]^2 + \|x\|^2\right)$ for a matrix A and a vector x together with the sublinear growth conditions on A, B, C and the Hölder inequality to get

$$\begin{aligned}
\left(\int_t^{t+h} \left\|A(s, X_s^i)\right\| ds \right)^p &\leq Kh^{p-1} \int_t^{t+h} \left(1 + \left\|X_s^i\right\|^p\right) ds \\
\left(\int_t^{t+h} \left\|B(s, X_s^i) \cdot \mu\right\| ds \right)^p &\leq Kh^{p-1} \int_t^{t+h} \left(1 + \left\|X_s^i\right\|^{2p}\right) ds + Kh \left\|\mu\right\|^{2p} . \\
\left(\int_t^{t+h} \left\|C(s, X_s^i) \cdot \phi^i\right\| ds \right)^p &\leq Kh^{p-1} \int_t^{t+h} \left(1 + \left\|X_s^i\right\|^{2p}\right) ds + Kh \left\|\phi^i\right\|^{2p} .
\end{aligned}$$

For $l \in \mathbb{N}$ we define $M_l = \sup_{0 \leq s \leq T} \mathbb{E}_\theta\left(\left\|X_s^i\right\|^l\right)$ and use the sublinear growth of Σ and the Burkholder-Davis-Gundy (BDG) inequality to get

$$\begin{aligned}
\mathbb{E}_\theta\left(\left\| \int_t^{t+h} \Sigma(s, X_s^i) dW_s^i \right\|^p \right) &\leq Kh^{\frac{p-2}{2}} \sum_{l,j=1}^r \mathbb{E}_{\theta_0}\left(\int_t^{t+h} \left|\Sigma_{lj}(s, X_s^i)\right|^p ds \right) \\
&\leq Kh^{\frac{p-2}{2}} \mathbb{E}_{\theta_0}\left(\int_t^{t+h} \left(1 + \left\|X_s^i\right\|^p\right) ds \right) \\
&\leq Kh^{\frac{p-2}{2}} h\left(1 + M_p\right) \leq Kh^{p/2}.
\end{aligned}$$

Taking expectations yields

$$\mathbb{E}_\theta\left(\left\|X_{t+h}^i - X_t^i\right\|^p\right) \leq Kh^{p/2}. \tag{3.10}$$

By assumption, the function $h = B'\Gamma^{-1}B$ is Lipschitz-continuous, such that

$$\begin{aligned}\mathbb{E}_{\theta_0}\left([\![V_{2i} - V_{2i}^n]\!]^p\right) &\leq \mathbb{E}_{\theta_0}\left(\sum_{k=0}^{n-1}\int_{t_k}^{t_{k+1}} [\![h(s, X_s^i) - h(t_k, X_{t_k}^i)]\!]^p ds\right) \\ &\leq K\sum_{k=0}^{n-1}\int_{t_k}^{t_{k+1}} \mathbb{E}_{\theta_0}\left(\left\|X_s^i - X_{t_k}^i\right\|^p\right) ds + K\frac{1}{n^p} \\ &\leq K\frac{1}{n^{p/2}}.\end{aligned}$$

The difference $U_{2i} - U_{2i}^n$ can be written as $U_{2i} - U_{2i}^n = A_1 + A_2 + A_3 + A_4$ with

$$\begin{aligned}A_1 &= \sum_{k=0}^{n-1}\int_{t_k}^{t_{k+1}} C(t_k, X_{t_k}^i)'\Gamma(t_k, X_{t_k}^i)^{-1}\left[A(t_k, X_{t_k}^i) - A(s, X_s^i)\right] ds \\ A_2 &= \sum_{k=0}^{n-1}\int_{t_k}^{t_{k+1}} \left[C(s, X_s^i)'\Gamma(s, X_s^i)^{-1} - C(t_k, X_{t_k}^i)'\Gamma(t_k, X_{t_k}^i)^{-1}\right] C(s, X_s^i) ds \cdot \phi^i \\ A_3 &= \sum_{k=0}^{n-1}\int_{t_k}^{t_{k+1}} \left[C(s, X_s^i)'\Gamma(s, X_s^i)^{-1} - C(t_k, X_{t_k}^i)'\Gamma(t_k, X_{t_k}^i)^{-1}\right] \Sigma(s, X_s^i) dW_s^i \\ A_4 &= \sum_{k=0}^{n-1}\int_{t_k}^{t_{k+1}} \left[C(s, X_s^i)'\Gamma(s, X_s^i)^{-1} - C(t_k, X_{t_k}^i)'\Gamma(t_k, X_{t_k}^i)^{-1}\right] B(s, X_s^i) ds \cdot \mu.\end{aligned}$$

For the first term we have

$$\begin{aligned}\|A_1\|^p &\leq \sum_{k=0}^{n-1}\left\|C(t_k, X_{t_k}^i)'\Gamma(t_k, X_{t_k}^i)^{-1}\right\|^p \int_{t_k}^{t_{k+1}} \left\|A(t_k, X_{t_k}^i) - A(s, X_s^i)\right\|^p ds \\ &\leq K\sum_{k=0}^{n-1}\left(1 + \left\|X_{t_k}^i\right\|^p\right)\left(\frac{1}{n^{p+1}} + \int_{t_k}^{t_{k+1}} \left\|X_{t_k}^i - X_s^i\right\|^p ds\right).\end{aligned}$$

Taking expectation, recalling the definition $M_l = \sup_{0\leq s\leq T}\mathbb{E}_\theta\left(\left\|X_s^i\right\|^l\right)$, and applying (3.10) and the Hölder inequality give

$$\begin{aligned}\mathbb{E}_\theta\left(\|A_1\|^p\right) &\leq \frac{K}{n^p}\left(1 + M_p\right) + \frac{K}{n^{p/2}} + K\sum_{k=0}^{n-1}\int_{t_k}^{t_{k+1}} \mathbb{E}_\theta\left(\left\|X_{t_k}^i\right\|^{2p}\right)^{1/2}\mathbb{E}_\theta\left(\left\|X_{t_k}^i - X_s^i\right\|^{2p}\right)^{1/2} ds \\ &\leq \frac{K}{n^p}\left[1 + M_p\right] + \frac{K}{n^{p/2}} + KM_{2p}^{1/2}\frac{1}{n^{p/2}} \\ &\leq \frac{K}{n^{p/2}}.\end{aligned}$$

For the second term we use Lipschitz-continuity of $C'\Gamma^{-1}$ and sublinear growth of C to get

$$\begin{aligned}[\![A_2]\!]^p &\leq \sum_{k=0}^{n-1}\int_{t_k}^{t_{k+1}} [\![C(s, X_s^i)'\Gamma(s, X_s^i)^{-1} - C(t_k, X_{t_k}^i)'\Gamma(t_k, X_{t_k}^i)^{-1}]\!]^p [\![C(s, X_s^i)]\!]^p ds \cdot \left\|\phi^i\right\|^p \\ &\leq K\left\|\phi^i\right\|^p \sum_{k=0}^{n-1}\int_{t_k}^{t_{k+1}}\left(\frac{1}{n^p} + \left\|X_s^i - X_{t_k}^i\right\|^p\left(1 + \left\|X_s^i\right\|^p\right) ds\right).\end{aligned}$$

We apply (3.10) and the general inequality $\mathbb{E}\left(\|X\|\,\|Y\|\,\|Z\|\right) \leq \mathbb{E}\left(\|X\|^2\right)^{1/2} \mathbb{E}\left(\|Y\|^4\right)^{1/4} \mathbb{E}\left(\|Z\|^4\right)^{1/4}$ yield

$$\begin{aligned}
\mathbb{E}_\theta\left([\![A_2]\!]^p\right) &\leq \frac{K}{n^p}\mathbb{E}_\theta\left(\left\|\phi^i\right\|^p\right) + \frac{K}{n^p}\mathbb{E}_\theta\left(\left\|\phi^i\right\|^{2p}\right)^{1/2}\sum_{k=0}^{n-1}\int_{t_k}^{t_{k+1}}\mathbb{E}_\theta\left(\left\|X_s^i\right\|^{2p}\right)^{1/2}ds \\
&\quad + K\mathbb{E}_\theta\left(\left\|\phi^i\right\|^{2p}\right)^{1/2}\sum_{k=0}^{n-1}\int_{t_k}^{t_{k+1}}\mathbb{E}_\theta\left(\left\|X_s^i - X_{t_k}^i\right\|^{4p}\right)^{1/4}\mathbb{E}_\theta\left(1+\left\|X_s^i\right\|^{4p}\right)^{1/4}ds \\
&\leq \frac{K}{n^p}\left[\mathbb{E}_\theta\left(\left\|\phi^i\right\|^p\right) + \mathbb{E}_\theta\left(\left\|\phi^i\right\|^{2p}\right)^{1/2}M_{2p}^{1/2}\right] + K\mathbb{E}_\theta\left(\left\|\phi^i\right\|^{2p}\right)^{1/2}\left(1+M_{4p}^{1/4}\right)\frac{1}{n^{p/2}} \\
&\leq \frac{K}{n^{p/2}}.
\end{aligned}$$

For the A_3-term we introduce $g_k(s) = \left[C(s,X_s^i)'\Gamma(s,X_s^i)^{-1} - C(t_k,X_{t_k}^i)'\Gamma(t_k,X_{t_k}^i)^{-1}\right]\Sigma(s,X_s^i)$ and the $(d\times r)$-valued process $H_s^n = \sum_{k=0}^{n-1}\mathbb{1}_{(t_k,t_{k+1}]}(s)g_k(s)$ and define the martingale $N_t^i = \int_0^t H_s^n dW_s^i$. Observe that we then have $A_3 = N_T^i$, such that the BDG and Hölder inequalities together with the Lipschitz-continuity of $C'\Gamma^{-1}$ and the growth condition on Σ give

$$\begin{aligned}
\mathbb{E}_{\theta_0}\left(\|A_3\|^p\right) = \mathbb{E}_{\theta_0}\left(\left\|N_T^i\right\|^p\right) &\leq K\sum_{k=0}^{n-1}\int_{t_k}^{t_{k+1}}\mathbb{E}_{\theta_0}\left([\![g_k(s)]\!]^p\right)ds \\
&\leq K\sum_{k=0}^{n-1}\int_{t_k}^{t_{k+1}}\mathbb{E}_{\theta_0}\left(\left\|X_s^i - X_{t_k}^i\right\|^p[\![\Sigma(s,X_s^i)]\!]^p\right)ds. \\
&\leq \frac{K}{n^{p/2}}.
\end{aligned}$$

The term A_4 satisfies

$$\|A_4\|^p \leq \sum_{k=0}^{n-1}\int_{t_k}^{t_{k+1}}[\![C(s,X_s^i)'\Gamma(s,X_s^i)^{-1} - C(t_k,X_{t_k}^i)'\Gamma(t_k,X_{t_k}^i)^{-1}]\!]^p[\![B(s,X_s^i)]\!]^p ds\,\|\mu\|^p$$

and can be treated as A_2 to get $\mathbb{E}_\theta\left(\|A_4\|^p\right) \leq \frac{K}{n^{p/2}}$.

□

Remark 3.13. A better result can often be obtained with a higher-order approximation. For the stochastic integral, this is obtained by using Ito's formula,

$$\int_{t_k}^{t_{k+1}} h(s,X_s^i)dX_s^i \approx H(t_{k+1},X_{k+1}^i) - H(t_k,X_k^i) - \frac{\Delta t}{2}\int_{t_k}^{t_{k+1}}\sum_{j,l=1}^{r}(H_{x_j,x_l}\Sigma_j\Sigma_l')(t_k,X_k^i),$$

with $h(t,x) = \nabla_x H(t,x)$. This, however, requires h to be of gradient-type, i.e., there should exist a differentiable function H such that h can be obtained as $h(t,x) = \nabla_x H(t,x)$. A higher-order approximation scheme is preferable, if the time step is not sufficiently small (non-high-frequency data) and/or the dynamics are highly non-linear.

Remark 3.14 (Likelihood discretization and Euler approximation)**.** The log-likelihood of the Euler-Maruyama approximation of the continuous-time model is proportional to

$$q^{\text{Euler}}(\mu,\varphi;X^{i,(n)}) \propto \sum_{k=0}^{n-1} F(t_k,X^i_{t_k},\mu,\varphi)'\Gamma(t_k,X^i_{t_k})^{-1}\Delta X^i_{t_k} - \frac{1}{2}\sum_{k=0}^{n-1} F(t_k,X^i_{t_k},\mu,\varphi)'\Gamma(t_k,X^i_{t_k})^{-1}F^i(t_k,X^i_{t_k},\mu,\varphi)\Delta t.$$

So the first-order discrete-time approximation of the conditional likelihood $q(\mu,\varphi;X^i)$ considered in Theorem 3.12 coincides (up to proportionality) with $q^{\text{Euler}}(\mu,\varphi;X^{i,(n)})$, such that the exact MLE of the approximating discrete-time Euler model coincides with the time-discretized MLE of the exact continuous-time model. This holds true more generally for arbitrary F (not necessarily affine in the mixed affects), if we choose as dominating measure measure for $\mathbb{Q}_{\mu,\varphi}$ the one that is induced by the diffusion with zero drift. If we call that measure $\mathbb{Q}_0$, the conditional density reads

$$\log\frac{d\mathbb{Q}_{\mu,\varphi}}{d\mathbb{Q}_0}(X^i) = \int_0^T F(s,X^i_s,\mu,\varphi)'\Gamma(s,X^i_s)^{-1}dX^i_s - \frac{1}{2}\int_0^T F(s,X^i_s,\mu,\varphi)'\Gamma(s,X^i_s)^{-1}F(s,X^i_s,\mu,\varphi)ds,$$

and the connection to the Euler-Maruyama log-likelihood is immediately visible. Since the MLE does not depend on the particular choice of the dominating measure, the MLE obtained from discretizing the continuous-time likelihood coincides with the one obtained from the Euler approximation.

Simulation study: Discretization bias for a non-Lipschitz drift

The derivations above assumed Lipschitz-continuity of the drift. A popular model which falls into the class of affine (in the parameter) drifts, but does not satisfy the Lipschitz requirements is the Fitzhugh-Nagumo (FHN) model (FitzHugh, 1955; Nagumo et al., 1962). We investigate how the discretization bias behaves as a function of the time discretization n^{-1} in this case. To provide some more background information, the FHN is a two-dimensional approximation of the well-known four-dimensional Hodgkin-Huxley neuronal model (Hodgkin and Huxley, 1952) and models the regenerative firing mechanism in an excitable neuron. However, since neural firing is a complex interplay of numerous cell processes, a stochastic extension which accounts for various unexplained noise sources is more suitable (Jensen et al., 2012),

$$\begin{aligned} dY_t &= \frac{1}{\varepsilon}\left(Y_t - Y_t^3 - Z_t + s\right)dt + \sigma_1 dW_{1,t}, \\ dZ_t &= \left(\gamma Y_t - Z_t + \eta\right)dt + \sigma_2 dW_{2,t}. \end{aligned} \tag{3.11}$$

The variable Y represents the membrane potential of a neuron, while the Z coordinate represents the recovery. The time scale separation ε is commonly $\ll 1$, such that Y lives on a much faster time scale than Z. The variable s is the input current. If $\gamma > 1$, the system has exactly one fixed point, which may be stable or unstable, depending on the specific parameter values. Under the reparametrization $\mu = (1/\varepsilon, s/\varepsilon, \gamma, \eta)'$, (3.11) can be written with drift as in Theorem 3.8. We assume to study a collection of N excitable neurons and model their membrane potentials Y^i via $dX_t^i = A(X_t^i) + C(X_t^i)(\mu + \phi^i)dt + \Sigma\, dW_t^i$, $0 \le t \le T$, where $X^i = (Y^i, Z^i)'$ and the ϕ^i are the i.i.d. $\mathcal{N}(0, \Omega)$-distributed random effects. Observe that despite being nonlinear in the state variable, the model equations are linear in the random effects and therefore an explicit likelihood is available. We assume here that both coordinates of X^i are observed. For all simulations, we let $\sigma_1 = 0.5, \sigma_2 = 0.3$ (assumed known), we fix $T = 20$ and choose the values of the unknown parameters as $\varepsilon = 0.1$, $s = 0.5$, $\gamma = 1.5$ and $\eta = 1.2$. With this choice of η the fixed point of the deterministic FHN system is stable, but small noise levels will suffice to induce large excursions through state space (spikes).

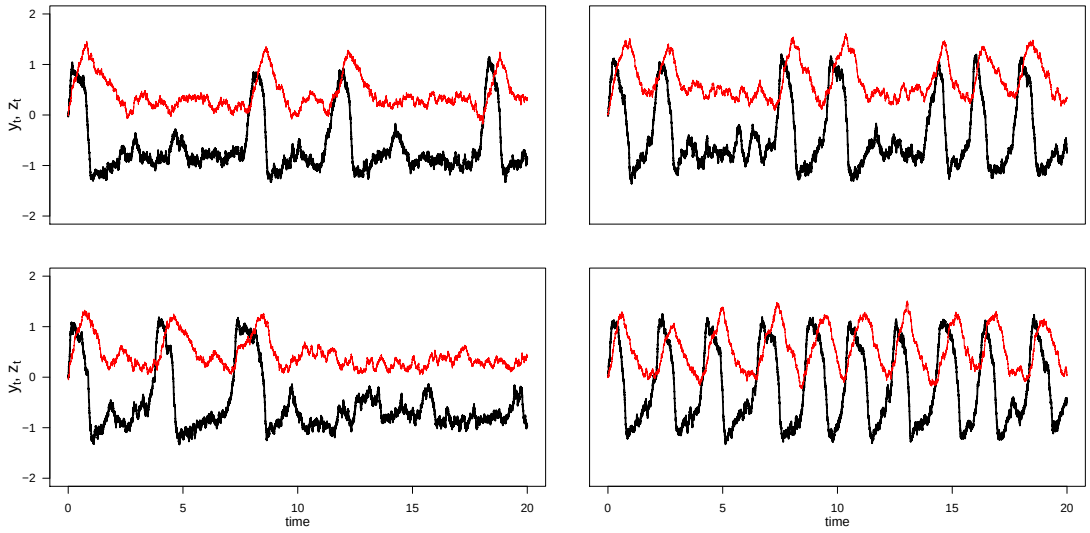


Figure 3.8: Trace plots of four realizations of the stochastic FHN model with random effects. The black traces correspond to the Y-coordinate, the red traces to the Z-coordinate. The corresponding realized parameter values of $\mu + \phi^i$ are (8.86,4.29,1.50,1.39), (10.16,7.49,1.73,1.40), (9.98,5.49,1.10,1.07) and (9.26,5.17, 1.84, 1.01).

The covariance matrix of the random effects is fixed as $\Omega = \operatorname{diag}\left(1.5^2, 1^2, 0.2^2, 0.2^2\right)$. The simulation settings are as follows. We simulate each trajectory $(X_t^1, \ldots, X_t^N)_{0 \le t \le T}$ with the Euler-Maruyama scheme and a simulation time step of $\delta = 10^{-4}$. The estimation is carried out on the

Table 3.2: FHN model. Shown are estimated relative bias and RMSE of $\hat{\mu}$ and $\mathrm{diag}(\hat{\Omega})$. The sample size is fixed to $N = 50$, but different sampling intervals are considered ($\Delta t = 0.001, 0.01, 0.1$). For each value of Δt, the estimation was repeated on $M = 500$ generated data sets.

		$\Delta t = 0.001$		$\Delta t = 0.01$		$\Delta t = 0.1$	
true value		rel. bias	RMSE	rel. bias	RMSE	rel. bias	RMSE
ε	0.10	0.003	0.022	0.030	0.037	0.356	0.356
s	0.50	0.001	0.033	0.002	0.033	0.009	0.035
γ	1.50	-0.000	0.031	-0.006	0.032	-0.079	0.121
η	1.20	-0.001	0.031	-0.005	0.032	-0.051	0.067
$1/\varepsilon$	10.00	-0.003	0.216	-0.028	0.349	-0.262	2.624
s/ε	5.00	-0.002	0.135	-0.026	0.188	-0.256	1.283
	2.25	-0.048	0.469	-0.155	0.539	-0.690	1.563
$\mathrm{diag}(\Omega)$	1.00	-0.025	0.197	-0.044	0.197	-0.281	0.318
	0.04	-0.035	0.010	-0.044	0.010	-0.212	0.012
	0.04	-0.007	0.010	-0.028	0.009	-0.218	0.012

thinned trajectory. We choose the sample size $N = 50$ and to illustrate how the discrete-time bias evolves, we repeat estimation for thinning factors $b = 10, 100, 1000$, which results in sampling intervals of $\Delta t = \delta \cdot b = 0.001, 0.01, 0.1$, respectively (note that the observation horizon is always fixed to $T = 20$). For all values of Δt, the estimation is repeated on $M = 500$ generated data sets. Figure 3.8 displays example trace plots of four realizations, which illustrate qualitatively different behaviors of the state process, depending on the realized values of the random effects. The black traces correspond to the Y-coordinate, the red traces to the Z-coordinate.

The estimation was done under the reparametrization $\mu = (1/\varepsilon, s/\varepsilon, \gamma, \eta)$. Estimates for the parameter ε and s on the original scale are obtained by transformation. Table 3.2 shows the sample estimates of the relative bias, $\frac{1}{M}\sum_{m=1}^{M} \frac{\hat{\theta}_j^{(m)} - \theta_j}{\theta_j}$, and of the root mean squared errors (RMSE), $\left(\frac{1}{M}\sum_{m=1}^{M}(\hat{\theta}_j^{(m)} - \theta_j)^2\right)^{1/2}$. Their estimation was based on samples with fixed sample size $N = 50$, but repeated for different sampling intervals Δt. The upper six rows in Table 3.2 show estimated bias and RSME for the fixed effects (on the original and on the transformed μ-scale) and the subsequent four rows correspond to results for the estimation of the diagonal of Ω. Despite the non-linearity of the model and the violation of standard assumptions on the diffusion drift for regularity of the model, the parameter estimation for high-frequency data and the chosen moderate sample size is very convincing (Table 3.2, first two columns), while still being satisfactory for observations

sampled at medium frequency (middle two columns). If observations are sampled at low frequency (last two columns in Table 3.2), the bias for the estimation of s, γ, η is still rather low (with 1%, 8% and 5% bias, respectively, as compared to the true parameter value). The estimation of ε is, however, highly biased. The variances of the random effects are all estimated with an error of about 21-28%, except for the variance of the random effect that adds to ε, which has an error of as high as 69%. This comes to no surprise, since non-linearity in the state requires denser observations. For ε we estimate the inverse of a small number, making the estimator unstable.

3.4.5 Asymptotic properties of the MLE for affine Gaussian SDMEMs

In this section we consider the setting of Theorem 3.8 and provide worked out proofs for consistency and asymptotic normality of the MLE $\hat{\theta}_N$. To maintain a fair degree of readability, we restrict ourselves to the case when all fixed effects are endowed with random effects, i.e., when $B(t,x) = C(t,x)$ in Theorem 3.8. In that case, we have $U_i := U_{1i} = U_{2i}$ and $V_i := V_{1i} = V_{2i} = Z_i$ and the likelihood (3.8) can be written as

$$p(\theta; X^i) = \frac{1}{\sqrt{\det(I_d + V_i\Omega)}} \mathrm{e}^{-\frac{1}{2}(\mu - V_i^{-1}U_i)' G^i(\Omega)(\mu - V_i^{-1}U_i) + \frac{1}{2}U_i' V_i^{-1} U_i},$$

and the Score function $S^i(\theta) = \left[\frac{d}{d\mu}\ell(\theta; X^i), \frac{d}{d\Omega}\ell(\theta; X^i)\right]$ in Theorem 3.10 simplifies to

$$\frac{d}{d\mu}\ell(\theta; X^i) = \gamma^i(\theta)' \qquad \text{and} \qquad \frac{d}{d\Omega}\ell(\theta; X^i) = -\frac{1}{2}\left[G^i(\Omega) - \gamma^i(\theta)\gamma^i(\theta)'\right],$$

with $\gamma^i(\theta) = G^i(\Omega)V_i^{-1}(U_i - V_i\mu)$.

In the proofs we follow a standard road. After verifying certain integrability requirements, we show asymptotic normality of the (normalized) score function[2]. The normalized negative log-likelihood is a contrast process and the Kullback-Leibler distance is the corresponding contrast function, which is shown to possess a unique minimum in the true parameter, which we may denote by θ_0. The consistency of the MLE follows from investigations of the continuity modulus. A Taylor expansion of the score function around the true parameter, together with its asymptotic normality and the consistency of the MLE, assures asymptotic normality of the MLE. In the sequel we assume that Θ is compact and convex and that the random matrix V_i is ($\mathbb{Q}_{0,0}^i$-a.s.) positive definite. In well-defined models this is not a restriction.

[2]We mean by asymptotic normality of a matrix A that its column-stacked linear transformation $\mathrm{vec}(A)$ is asymptotically normal.

Lemma 3.15 (Moment properties)**.** *For all $\theta \in \Theta$ the following statements hold.*

(i) For all $u \in \mathbb{R}^d$, $\mathbb{E}_\theta\left(e^{u'(I_d+V_i\Omega)^{-1}U_i}\right) < \infty$.

(ii) The score function is centered, $\mathbb{E}_\theta\left(\frac{d}{d\mu}\ell^i(\theta;X^i)\right) = 0$ and $\mathbb{E}_\theta\left(\frac{d}{d\Omega}\ell^i(\theta;X^i)\right) = 0$.

(iii) The score function has finite second moments.

Proof of Lemma 3.15. We begin with (i). First of all we recall that $G^i(\Omega)$ and $(\Omega+\Omega V_i\Omega)^{-1}$ are positive definite, such that the identity $G^i(\Omega) = \Omega^{-1} - (\Omega+\Omega V_i\Omega)^{-1}$ assures the bound $u'G^i(\Omega)u \leq u'\Omega^{-1}u$ for any $u \in \mathbb{R}^d$. Let us now fix $u \in \mathbb{R}^d$ and define $\tau = (\mu+u,\Omega)$. The likelihoods then satisfy the identity $p(\theta;X^i)\mathrm{e}^{u'\gamma^i(\theta)} = p(\tau;X^i)\mathrm{e}^{\frac{1}{2}u'G^i(\Omega)u}$, such that

$$\mathbb{E}_\theta\left(\mathrm{e}^{u'\gamma^i(\theta)}\right) = \mathbb{E}_{0,0}\left(\mathrm{e}^{u'\gamma_i(\theta)}p(\theta;X^i)\right) = \mathbb{E}_{0,0}\left(\mathrm{e}^{\frac{1}{2}u'G^i(\Omega)u}p(\tau;X^i)\right) = \mathbb{E}_\tau\left(\mathrm{e}^{\frac{1}{2}u'G^i(\Omega)u}\right) \leq \mathrm{e}^{\frac{1}{2}u'\Omega^{-1}u},$$

which is finite. Note that for arbitrary $u \in \mathbb{R}^d$ we further have the bound

$$u'G^i(\Omega)\mu \leq u'G^i(\Omega)\mu + \frac{1}{2}u'G^i(\Omega)u + \frac{1}{2}\mu'G^i(\Omega)\mu = \frac{1}{2}(u+\mu)'G^i(\Omega)(u+\mu) \leq \frac{1}{2}(u+\mu)'\Omega^{-1}(u+\mu),$$

and therefore,

$$\begin{aligned}\mathbb{E}_\theta\left(\mathrm{e}^{u'(I_d+V_i\Omega)^{-1}U_i}\right) &= \mathbb{E}_\theta\left(\mathrm{e}^{u'\gamma^i(\theta)}\mathrm{e}^{u'(I_d+V_i\Omega)^{-1}V_i\mu}\right) = \mathbb{E}_\theta\left(\mathrm{e}^{u'\gamma^i(\theta)}\mathrm{e}^{u'G^i(\Omega)\mu}\right)\\ &\leq \mathbb{E}_\theta\left(\mathrm{e}^{u'\gamma^i(\theta)}\right)\mathrm{e}^{\frac{1}{2}(u+\mu)'\Omega^{-1}(u+\mu)} < \infty,\end{aligned}$$

completing the proof of (i).

For (ii) we restrict ourselves to only showing the first identity (the second identity in (ii) can be shown in the same spirit, but with choosing $\tau = (\mu,\Omega_0)$ below instead). In general, for any $\tau = (\mu_0,\Omega)$ we have

$$\left[\frac{d}{d\mu}\ell(\theta;X^i)\right]\frac{p(\theta;X^i)}{p(\tau;X^i)} = \frac{\frac{d}{d\mu}p(\theta;X^i)}{p(\theta;X^i)}\frac{p(\theta;X^i)}{p(\tau;X^i)} = \frac{\frac{d}{d\mu}p(\theta;X^i)}{p(\tau;X^i)} = \frac{d}{d\mu}\left[\frac{p(\theta;X^i)}{p(\tau;X^i)}\right].$$

Taking expectations gives

$$\mathbb{E}_\theta\left(\frac{d}{d\mu}\ell(\theta;X^i)\right) = \mathbb{E}_\tau\left(\left[\frac{d}{d\mu}\ell(\theta;X^i)\right]\frac{p(\theta;X^i)}{p(\tau;X^i)}\right) = \mathbb{E}_\tau\left(\frac{d}{d\mu}\left[\frac{p(\theta;X^i)}{p(\tau;X^i)}\right]\right).$$

If expectation and integration can be interchanged, the claim follows. The validity of interchanging these two limit operations is shown below. For convenience, we use $\mu_0 = 0$ and introduce the shorthand notation

$$\alpha^i(\theta) = \frac{p^i(\theta;X^i)}{p^i(\tau;X^i)} = \frac{d\mathbb{Q}^i_\theta}{d\mathbb{Q}^i_\tau} = \mathrm{e}^{\mu'G^i(\Omega){V_i}^{-1}U_i - \frac{1}{2}\mu'G^i(\Omega)\mu}.$$

Let $\mu^*, h \in \mathbb{R}^d$, and $(\epsilon_n)_{n\in\mathbb{N}}$ a zero sequence in $\mathbb{R}_{>0}$, and define $\theta^* = (\mu^*, \Omega)$ and $\theta_n^* = (\mu^* + \epsilon_n h, \Omega)$. We study the function $f_n(X^i) = \frac{\alpha^i(\theta_n^*) - \alpha^i(\theta^*)}{\|h\|\epsilon_n}$. Then, as n goes to infinity, $f_n(X^i) \to \frac{d}{d\mu}\alpha^i(\theta)\big|_{\theta=\theta^*}$. To prove that this pointwise convergence implies the convergence of the integrals,

$$\lim_{n\to\infty} \int_{C_T} f_n(X^i) d\mathbb{Q}_\tau^i(X^i) = \int_{C_T} \lim_{n\to\infty} f_n(X^i) d\mathbb{Q}_\tau^i(X^i),$$

we show the existence of a non-negative, $\mathbb{Q}_\tau^i$-integrable random variable $M(\Omega) : (C_T, \mathcal{C}_T) \to (\mathbb{R}, \mathcal{B})$, such that $\left\| \frac{d}{d\mu}\alpha^i(\theta)|_{\theta=\theta^*} \right\| \leq M(\Omega)$ $\mathbb{Q}_\tau^i$-a.s. Without loss of generality assume that there is $R > 0$ such that for all $n \in \mathbb{N}$, the vectors $\mu^* - \epsilon_n h$ and $\mu^* + \epsilon_n h$ are contained in $\overline{B}_d(R)$, the closure of the ball in $\mathbb{R}^d$ with center 0 and radius R. By the mean value theorem there is for each n a $c_n \in [0,1]$ such that $|f_n| = \left\| \frac{\partial}{\partial\mu}\alpha^i(\eta_{n,h}^*) \right\|$ with $\eta_{n,h}^* = \left(\mu_{n,h}^*, \Omega\right)$ and $\mu_{n,h}^* = (1 - c_n)(\mu^* + \epsilon_n h) + c_n \mu^*$. In particular, due to the convexity of $\overline{B}_d(R)$, we have $\mu_{n,h}^* \in \overline{B}_d(R)$ for all n and thus

$$|f_n| \leq \sup_{n\in\mathbb{N}} \left\| \frac{\partial}{\partial\mu}\alpha^i(\eta_{n,h}^*) \right\| \leq \sup_{\mu\in\overline{B}_d(R)} \left\| \frac{\partial}{\partial\mu}\alpha^i(\theta) \right\| = \sup_{\mu\in\overline{B}_d(R)} \left\| \gamma^i(\theta)' \alpha^i(\theta) \right\|.$$

In the proof of (i) we showed that $G^i(\Omega) - \Omega^{-1}$ is negative semi-definite and therefore

$$\begin{aligned} \left\|\gamma^i(\theta)\right\| &\leq \left\| (I_d + V_i\Omega)^{-1} U_i \right\| + \left\| G^i(\Omega)\mu \right\| \leq \left\| (I_d + V_i\Omega)^{-1} U_i \right\| + [\![\Omega^{-1}]\!] \|\mu\| \\ &\leq \left\| (I_d + V_i\Omega)^{-1} U_i \right\| + R\,[\![\Omega^{-1}]\!] \end{aligned}$$

and since $G^i(\Omega)$ is positive definite, $\alpha^i(\theta) \leq \mathrm{e}^{\mu' G^i(\Omega) V_i^{-1} U_i} = \mathrm{e}^{\mu'(I_d + V_i\Omega)^{-1} U_i}$. For any $u \in \overline{B}_d(R), z \in \mathbb{R}^d$ one has $\mathrm{e}^{u'z} \leq \left[\mathrm{e}^{Rz_i} + \mathrm{e}^{-Rz_i}\right] \mathrm{e}^{\sum_{j=2}^d u_j z_j}$, which generalizes to $\mathrm{e}^{u'z} \leq \prod_{j=1}^d \left[\mathrm{e}^{Rz_j} + \mathrm{e}^{-Rz_j}\right] = \sum_{\beta\in\mathcal{K}} \mathrm{e}^{R\beta' z}$, with $\mathcal{K} = \{u \in \mathbb{R}^d : u_j \in \{-1, 1\}, j = 1, \ldots, d\}$. Therefore, for any $\mu \in \overline{B}_d(R)$,

$$\alpha^i(\theta) \leq \mathrm{e}^{\mu' G^i(\Omega) V_i^{-1} U_i} \leq \sum_{\beta\in\mathcal{K}} \mathrm{e}^{R\beta'(I_d + V_i\Omega)^{-1} U_i} =: \tilde{M}(\Omega).$$

such that $|f_n|$ can be bounded by

$$\begin{aligned} |f_n| &\leq \sup_{\mu\in\overline{B}_d(R)} \left(\left\| (I_d + V_i\Omega)^{-1} U_i \right\| + R[\![\Omega^{-1}]\!] \right) \tilde{M}(\Omega) = \left(\left\| (I_d + V_i\Omega)^{-1} U_i \right\| + R[\![\Omega^{-1}]\!] \right) \tilde{M}(\Omega) \\ &=: M(\Omega), \end{aligned}$$

which is $\mathbb{Q}_\tau^i$-integrable by (i), since $\mathbb{E}_\tau \left((I_d + V_i\Omega)^{-1} U_i \cdot \mathrm{e}^{u'(I_d + V_i\Omega)^{-1} U_i} \right) < \infty$ for any $u \in \mathbb{R}^d$. Thus, integration and differentiation can be interchanged. For the proof of (iii) we refer to section 2.3.1. in Paper III.

□

The asymptotic normality of the normalized score function is now a consequence of the law of large numbers (LLN), together with the standard multivariate central limit theorem (CLT).

Theorem 3.16 (Asymptotic normality of the normalized score function)**.**
Let $\mathcal{S}^i(\theta; X^i) = \left(\frac{d}{d\mu}\ell^i(\theta; X^i), vec\left[\frac{d}{d\Omega}\ell^i(\theta; X^i)\right]'\right)'$ *and* $\mathcal{S}_N(\theta; \boldsymbol{X}) = \sum_{i=1}^N \mathcal{S}^i(\theta; X^i)$ *denote the vectorized score function. For all* $\theta \in \Theta$*, under* $\mathbb{Q}_\theta$ *and as* N *tends to infinity, we have*

$$\frac{1}{\sqrt{N}}\mathcal{S}_N(\theta; \boldsymbol{X}) \to \mathcal{N}(0, \mathcal{I}(\theta))$$

where $\mathcal{I}(\theta) = \mathbb{C}ov_\theta\left(\mathcal{S}^i(\theta; X^i)\right)$.

To show the consistency and asymptotic normality of the MLE $\hat{\theta}_N$, we distinguish two cases, depending on whether $B(t,x)'\Gamma(t,x)^{-1}B(t,x)$ is constant as a function of x. We let $m = d + d^2$, denote by $\mathbb{\lambda}^m$ the Lebesgue measure on $\mathbb{R}^m$ and identify $\mathbb{R}^{d\times d}$ with $\mathbb{R}^{d^2}$ (by considering the vectorized versions of matrices in $\mathbb{R}^{d\times d}$). We make the following additional assumptions.

(C) (i) The function $B(t,x)'\Gamma(t,x)^{-1}B(t,x)$ is not constant in x, and under $\mathbb{Q}^i_{0,0}$, the (vectorized) $\mathbb{R}^m$-valued random vector (U_i, V_i) admits a continuous density function $f(u,v)$ (w.r.t. the Lebesgue measure $\mathbb{\lambda}^m$), which is positive on an open ball of $\mathbb{R}^m \supset \Theta$.

(ii) We assume that $\theta \in \Theta$ can be bounded via: $\|\mu\| \leq M_1$ and $M_{2,1} \leq [\![\Omega]\!] \leq M_{2,2}$, where $M_1, M_{2,1}, M_{2,2}$ are positive constants (which exists because Θ is bounded).

(ii) The true value θ_0 belongs to $\text{int}(\Theta)$ and the matrix $\mathcal{I}(\theta_0)$ is invertible.

Theorem 3.17 (Continuity of KL information and uniqueness of its minimum)**.**
Under (C) the Kullback-Leibler (KL) information of $\mathbb{Q}_{\theta_0}$ *w.r.t.* $\mathbb{Q}_\theta$*, denoted by* $K(\mathbb{Q}_{\theta_0}, \mathbb{Q}_\theta)$*, is continuous on* $\mathbb{R}^m$ *and has a unique minimum at* $\theta = \theta_0$*.*

Proof of Theorem 3.17. Recall that in general $KL(\mathbb{Q}_{\theta_0}, \mathbb{Q}_\theta) \geq 0$ and equality holds if and only if $\mathbb{Q}_{\theta_0} = \mathbb{Q}_\theta$. To show the uniqueness of the minimum at θ_0 we show that $\mathbb{Q}_{\theta_0} = \mathbb{Q}_\theta$ implies $\theta_0 = \theta$. Note that $p(\theta; X^i)$ can be written as a function of only the sufficient statistics and with slight abuse of notation we write therefore $p(\theta; X^i) = p(\theta; U_i, V_i)$. By (C)(i), the distribution $\mathbb{Q}_{0,0}^{(U_i,V_i)}$ of (U_i, V_i) under $\mathbb{Q}_{0,0}$ admits the $\mathbb{\lambda}^m$-density $\frac{d\mathbb{Q}_{0,0}^{(U_i,V_i)}}{d\mathbb{\lambda}^m}(u,v) = f(u,v)$ and by Lemma A.3 we have

$\mathbb{Q}_\theta^{(U_i,V_i)} \ll \mathbb{Q}_{0,0}^{(U_i,V_i)}$ with density

$$\frac{d\mathbb{Q}_\theta^{(U_i,V_i)}}{d\mathbb{Q}_{0,0}^{(U_i,V_i)}}(u,v) = \mathbb{E}_{0,0}\left[\frac{d\mathbb{Q}_\theta}{d\mathbb{Q}_{0,0}}|(U_i,V_i)=(u,v)\right] = \mathbb{E}_{0,0}\left[p(\theta;U_i,V_i)|(U_i,V_i)=(u,v)\right] = p(\theta;u,v).$$

Hence, the distribution $\mathbb{Q}_\theta^{(U_i,V_i)}$ of the random vector (U_i,V_i) under $\mathbb{Q}_\theta$ has λ^m-density

$$f_\theta(u,v) = \frac{d\mathbb{Q}_\theta^{(U_i,V_i)}}{d\lambda^m}(u,v) = \frac{d\mathbb{Q}_\theta^{(U_i,V_i)}}{d\mathbb{Q}_{0,0}^{(U_i,V_i)}}(u,v)\frac{d\mathbb{Q}_{0,0}^{(U_i,V_i)}}{d\lambda^m}(u,v) = p(\theta;u,v)f(u,v).$$

Equality of the distributions $\mathbb{Q}_\theta$, $\mathbb{Q}_{\theta_0}$ implies $\mathbb{Q}_\theta^{(U_i,V_i)} = \mathbb{Q}_{\theta_0}^{(U_i,V_i)}$ and due to the (λ^m-a.s.) uniqueness of the Radon-Nikodym densities f_θ, f_{θ_0}, one therefore has $f_\theta = f_{\theta_0}$ λ^m-almost everywhere. The continuity of f_θ and f_{θ_0} assures that equality even holds everywhere, i.e. $f_\theta = f_{\theta_0}$. By assumption, $f(u,v)$ is positive on an open ball B of $\mathbb{R}^m$. Let $F_B = B \cap (\mathbb{R}^d \times \mathcal{S}_d(\mathbb{R}))$ and $G_B = B \cap (\mathbb{R}^d \times \mathcal{S}_d(\mathbb{R}))^c$. Here $\mathcal{S}_d(\mathbb{R})$ is the set of symmetric, positive definite $(d \times d)$-matrices and the superscript c denotes the set complement, such that $B = F_B \cup G_B$. Then F_B is open and as such $\lambda^m(F_B) > 0$. On G_B we have that $p(\theta;u,v) = p(\theta_0;u,v) = 0$ (which does not help us for deducing equality of θ and θ_0), whereas on F_B it holds that $p(\theta;u,v) = p(\theta_0;u,v) > 0$ or, equivalently, that (note that for $(u,v) \in F_B$ the inverse v^{-1} exists)

$$\left(\frac{\det(I_d + v\Omega_0)}{\det(I_d + v\Omega)}\right)^{1/2} = \mathrm{e}^{-\frac{1}{2}(\mu_0 - v^{-1}u)'G^i(\Omega_0)((\mu_0 - v^{-1}u) + \frac{1}{2}(\mu - v^{-1}u)'G^i(\Omega)(\mu - v^{-1}u)}.$$

This implies $\Omega = \Omega_0$ and $\mu = \mu_0$, i.e., $\theta = \theta_0$. To establish continuity of $\theta \mapsto K(\mathbb{Q}_{\theta_0}, \mathbb{Q}_\theta)$, we recall that

$$\begin{aligned} K(\mathbb{Q}_{\theta_0}, \mathbb{Q}_\theta) &= \int_{\mathbb{R}^m} f_{\theta_0}(u,v) \log \frac{f_{\theta_0}(u,v)}{f_\theta(u,v)} d\lambda(u,v) = \int_{\mathbb{R}^m} f_{\theta_0}(u,v)\left[\ell(\theta_0;u,v) - \ell(\theta;u,v)\right] d\lambda(u,v) \\ &= \mathbb{E}_{\theta_0}\left(h(\theta;U_i,V_i)\right), \end{aligned}$$

with $h(\theta;u,v) = \ell(\theta_0;u,v) - \ell(\theta;u,v)$. Obviously, $\mu \mapsto h(\theta;U_i,V_i)$ is $\mathbb{Q}_{\theta_0}^{(U_i,V_i)}$-a.s. continuous. Continuity of matrix multiplication and addition, of taking the inverse, of the determinant and of the logarithm assure a.s. continuity of $\Omega \mapsto h(\theta;U_i,V_i)$. Thus, $\theta \mapsto h(\theta;U_i,V_i)$ is (a.s.) continuous. The continuity of $\theta \mapsto K(\mathbb{Q}_{\theta_0}, \mathbb{Q}_\theta)$ follows from dominated convergence. That the limit operations in fact may be interchanged is justified by the fact that there is a $\mathbb{Q}_{\theta_0}$-integrable random variable M such that $h(\theta;x) \le M(x)$ for all $x \in C_T$. We verify the existence of M below. We can write $2h(\theta;U_i,V_i)$ as the sum $A_1 + \ldots + A_5$, with

$$\begin{aligned} A_1 &= \log\left(\frac{\det(I_d + V_i\Omega)}{\det(I_d + V_i\Omega_0)}\right) & A_2 &= U_i'V_i^{-1}\left[G^i(\Omega) - G^i(\Omega_0)\right]V_i^{-1}U_i \\ A_3 &= \mu' G^i(\Omega)\mu, & A_4 &= \mu' G^i(\Omega)V_i^{-1}U_i \\ A_5 &= \left[\mu_0' G^i(\Omega_0)\mu_0 - \mu_0' G^i(\Omega_0)V_i^{-1}U_i\right]. \end{aligned}$$

A_5 does not depend on θ and its $\mathbb{Q}_{\theta_0}$-integrability is straightforward by Theorem 3.15(i). With arguments as used previously we can bound A_3 by $\|A_3\| \leq \|\mu\|^2 [\![G^i(\Omega)]\!] \leq \frac{M_1^2}{M_{2,1}}$. For A_4 we note that

$$\begin{aligned} G^i(\Omega)V_i^{-1}U_i &= (I_d + V_i\Omega)^{-1}U_i = (I_d + V_i\Omega)^{-1}(I_d + V_i\Omega - V_i\Omega + V_i\Omega_0)(I_d + V_i\Omega_0)^{-1}U_i \\ &= \left[I_d + G^i(\Omega)\left(\Omega_0 - \Omega\right)\right](I_d + V_i\Omega_0)^{-1}U_i. \end{aligned}$$

This, together with $[\![I_d + G^i(\Omega)(\Omega_0 - \Omega)]\!] \leq C\frac{M_{2,2}}{M_{2,1}}$, yields

$$|A_4| \leq \|\mu\| \left\|G^i(\Omega)V_i^{-1}U_i\right\| \leq CM_1 \left\|(I_d + V_i\Omega_0)^{-1}U_i\right\| \leq C \left\|(I_d + V_i\Omega_0)^{-1}U_i\right\|,$$

which is $\mathbb{Q}_{\theta_0}$-integrable by Theorem 3.15(i). For A_2 we observe that

$$A_2 = \left[(I_d + V_i\Omega_0)^{-1}U_i\right]' \underbrace{(I_d + V_i\Omega_0)\, V_i^{-1}\left[(I_d + V_i\Omega)^{-1} - (I_d + V_i\Omega_0)^{-1}\right](I_d + V_i\Omega_0)}_{=:\tilde{A}_2} \left[(I_d + V_i\Omega_0)^{-1}U_i\right],$$

such that $|A_2| \leq \left\|(I_d + V_i\Omega_0)^{-1}U_i\right\|^2 [\![\tilde{A}_2]\!]$. For $\tilde{A}_2$ we have

$$\begin{aligned} \tilde{A}_2 &= (I_d + V_i\Omega_0)\, V_i^{-1}\left[(I_d + V_i\Omega)^{-1} - (I_d + V_i\Omega_0)^{-1}\right](I_d + V_i\Omega_0) \\ &= (I_d + V_i\Omega_0)\, V_i^{-1}\left[(I_d + V_i\Omega)^{-1}(I_d + V_i\Omega_0) - (I_d + V_i\Omega)^{-1}(I_d + V_i\Omega)\right] \\ &= (I_d + V_i\Omega_0)\, V_i^{-1}(I_d + V_i\Omega)^{-1}\left[(I_d + V_i\Omega_0) - (I_d + V_i\Omega)\right] \\ &= (I_d + V_i\Omega_0)\, V_i^{-1}(I_d + V_i\Omega)^{-1}V_i\left(\Omega - \Omega_0\right), \end{aligned}$$

such that (K is a generic constant and may vary from line to line)

$$\begin{aligned} [\![\tilde{A}_2]\!] &\leq [\![I_d + V_i\Omega_0]\!][\![V_i]\!]^{-1}[\![I_d + V\Omega]\!][\![V_i]\!][\![\Omega_0 - \Omega]\!] = \frac{[\![I_d + V_i\Omega_0]\!]}{[\![I_d + V_i\Omega]\!]}[\![\Omega_0 - \Omega]\!] \\ &\leq \frac{K}{[\![I_d + V_i\Omega]\!]} + \frac{[\![V_i\Omega_0]\!]}{[\![I_d + V_i\Omega]\!]} + 2M_{2,2} \leq K \cdot 1 + M_{2,2}\frac{[\![V_i]\!]}{[\![I_d + V_i\Omega]\!]} + 2M_{2,2} \\ &\leq K + \frac{M_{2,2}}{M_{2,1}}\frac{[\![V_i\Omega]\!]}{[\![I_d + V_i\Omega]\!]} \leq K + \frac{M_{2,2}}{M_{2,1}}. \end{aligned}$$

Therefore we get as a bound

$$|A_2| \leq \left\|(I_d + V_i\Omega_0)^{-1}U_i\right\|^2 [\![\tilde{A}_2]\!] \leq K \left\|(I_d + V_i\Omega_0)^{-1}U_i\right\|^2,$$

which is $\mathbb{Q}^i_{\theta_0}$-integrable by Theorem 3.15(i). It remains to check the first term A_1. For this, note that for a $(d \times d)$- matrix A we have $\det(A) = \prod_{i=1}^d \lambda_i(A)$, where $\lambda_i(A)$ are the eigenvalues of A. Since the spectral norm of A is $[\![A]\!]_S = \max_{i=1,\ldots,d} |\lambda_i(A)|$, one has $\det(A) \leq [\![A]\!]_S^d \leq C[\![A]\!]^d$. We can bound the argument of the logarithm from above by

$$\frac{\det(I_d + V_i\Omega)}{\det(I_d + V_i\Omega_0)} = \det\left((I_d + V_i\Omega)(I_d + V_i\Omega_0)^{-1}\right) \leq K[\![I_d + V_i\Omega]\!]^d[\![(I_d + V_i\Omega_0)^{-1}]\!]^d \leq K\left(\frac{[\![I_d + V_i\Omega]\!]}{[\![I_d + V_i\Omega_0]\!]}\right)^d.$$

As before, this random quantity can itself be bounded by a constant,

$$\frac{[\![I_d + V_i\Omega]\!]}{[\![I_d + V_i\Omega_0]\!]} \leq K\frac{1}{[\![I_d + V_i\Omega_0]\!]} + \frac{[\![V_i\Omega_0\Omega_0^{-1}]\!][\![\Omega]\!]}{[\![I_d + V_i\Omega_0]\!]} \leq K + \frac{1}{\|I_d + V_i\Omega_0\|} \leq K,$$

such that the monotonicity of log implies $A_1 = \log\left(\frac{\det(I_d+V_i\Omega)}{\det(I_d+V_i\Omega_0)}\right) \leq \log K$. The term $-A_1$ is treated analogously, such that we conclude $|A_1| \leq K$. We have hence shown that $2\,|h(\theta; U_i, V_i)| \leq M$ and M is integrable with respect to $\mathbb{Q}_{\theta_0}$. This completes the continuity proof.

□

Theorem 3.18 (Weak consistency and asymptotic normality of the MLE)**.** *Assume (C) and let $\hat{\theta}_N$ be an MLE defined as any solution of $\ell_N(\hat{\theta}_N; \boldsymbol{X}) = \sup_{\theta\in\Theta} \ell_N(\theta; \boldsymbol{X})$. Then the following assertions hold.*

(i) $\hat{\theta}_N$ converges in $\mathbb{Q}_{\theta_0}$- probability to θ_0.

(ii) As N tends to infinity, $\sqrt{N}vec\left(\hat{\theta}_N - \theta_0\right) \to \mathcal{N}\left(0, \mathcal{I}^{-1}(\theta_0)\right)$ under $\mathbb{Q}_{\theta_0}$.

Proof of Theorem 3.18. We start by showing (i), the weak consistency of the MLE. Recall that $\ell_N(\theta; \boldsymbol{X})$ is the N-sample log-likelihood and that the $\ell(\theta; X^i), i = 1, \ldots, N$, are i.i.d. random variables with $\mathbb{E}_{\theta_0}\left(\ell(\theta_0; X^i) - \ell(\theta; X^i)\right) = K(\mathbb{Q}_{\theta_0}, \mathbb{Q}_\theta)$. The LLN therefore implies

$$\frac{1}{N}\left(\ell_N(\theta_0; \boldsymbol{X}) - \ell_N(\theta; \boldsymbol{X})\right) \xrightarrow{\mathbb{Q}_{\theta_0}} K(\mathbb{Q}_{\theta_0}, \mathbb{Q}_\theta).$$

Hence, $K_N(\theta) = -\frac{1}{N}\ell_N(\theta; \boldsymbol{X})$ is a contrast process with contrast function $\theta \mapsto K(\mathbb{Q}_{\theta_0}, \mathbb{Q}_\theta)$. It therefore remains to study the continuity modulus of $K_N(\theta)$, given by

$$\omega_N(\eta) = \sup_{\|\theta-\theta^*\|\leq\eta, \theta^*\in\Theta} |K_N(\theta) - K_N(\theta^*)| = \sup_{\|\theta-\theta^*\|\leq\eta, \theta^*\in\Theta} \frac{1}{N}\left|\ell_N(\theta; \boldsymbol{X}) - \ell_N(\theta^*; \boldsymbol{X})\right|.$$

The mean value theorem and the convexity of Θ assure

$$|\ell_N(\theta; \boldsymbol{X}) - \ell_N(\theta_0; \boldsymbol{X})| \leq \int_0^1 \|\mathcal{S}_N\left(\theta_0 + t(\theta - \theta_0); \boldsymbol{X}\right)\| \, dt \cdot \|\theta - \theta_0\| \leq \sup_{\tilde{\theta}\in\Theta} \left\|\mathcal{S}_N(\tilde{\theta}; \boldsymbol{X})\right\| \|\theta - \theta_0\|$$

and therefore $\omega_N(\eta) \leq K\eta\frac{1}{N}\sup_{\tilde{\theta}\in\Theta}\left\|\mathcal{S}_N(\tilde{\theta}; \boldsymbol{X})\right\|$. We now verify that the right hand side of this inequality has finite expectation under $\mathbb{Q}_{\theta_0}$. To this end, we will show the existence of a $\mathbb{Q}_{\theta_0}$-integrable random variable $M(\theta_0)$ such that $\|\mathcal{S}_N(\theta; \boldsymbol{X})\| \leq M(\theta_0)$ a.s. for any $\theta \in \Theta$. Obviously, the score function can be bounded by

$$\|\mathcal{S}_N(\theta; \boldsymbol{X})\| \leq \sum_{i=1}^N \left(\left\|\frac{d}{d\mu}\ell(\theta; X^i)\right\| + \left[\!\left[\frac{d}{d\Omega}\ell(\theta; X^i)\right]\!\right]\right) \leq K\sum_{i=1}^N \left(\|\gamma^i(\theta)\| + \|\gamma^i(\theta)\|^2 + [\![G^i(\Omega)]\!]\right).$$

In the following, we find bounds on $\left\|\gamma^i(\theta)\right\|$ and $[\![G^i(\Omega)]\!]$ that are uniform in $\theta \in \Theta$. We start by noting the identities

$$\begin{aligned}
\gamma^i(\theta) &= (I_d + V_i\Omega)^{-1}U_i - G^i(\Omega)\mu, \\
(I_d + V_i\Omega)^{-1}U_i &= \left[(I_d + V_i\Omega)^{-1}(I_d + V_i\Omega_0)\right](I_d + V_i\Omega_0)^{-1}U_i, \qquad (3.12)\\
(I_d + V_i\Omega)^{-1}(I_d + V_i\Omega_0) &= (I_d + V_i\Omega)^{-1}[I_d + V_i\Omega - V_i\Omega_0 + V_i\Omega] = I_d + G^i(\Omega)(\Omega_0 - \Omega),
\end{aligned}$$

which imply $(I_d + V_i\Omega)^{-1}U_i = \left[I_d + G^i(\Omega)(\Omega_0 - \Omega)\right]B^i(\Omega_0)$, with $B^i(\Omega_0) := (I_d + V_i\Omega_0)^{-1}U_i$. Moreover, we have $[\![G^i(\Omega)]\!] \leq K[\![\Omega^{-1}]\!] \leq K\frac{1}{M_{2,1}}$. Combining this bound with (3.12), we get

$$\begin{aligned}
\left\|\gamma^i(\theta)\right\| &\leq [\![I_d + G^i(\Omega)(\Omega_0 - \Omega)]\!][\![B_i(\Omega_0)]\!] + K\frac{M_{2,2}}{M_{2,1}} \leq K\left(1 + [\![\Omega]\!]^{-1}([\![\Omega_0]\!] + [\![\Omega]\!])\right)[\![B^i(\Omega_0)]\!] \\
&\leq K\left(1 + [\![B^i(\Omega_0)]\!]\right) =: \kappa_i(\Omega_0),
\end{aligned}$$

and conclude, with $M(\theta_0) = \sum_{i=1}^N M^i(\theta_0)$ and $M^i(\theta_0) = \left(1 + \kappa_i(\Omega_0) + \kappa_i(\Omega_0)^2\right)$, that $\sup_{\theta\in\Theta}\left\|\frac{d}{d\theta}\ell_N(\theta; \boldsymbol{X})\right\| \leq KM(\theta_0)$ and therefore

$$\mathbb{E}_{\theta_0}\left(\sup_{\theta\in\Theta}\left\|\frac{d}{d\theta}\ell_N(\theta; \boldsymbol{X})\right\|\right) \leq K\mathbb{E}_{\theta_0}\left(M(\theta_0)\right) = KN\mathbb{E}_{\theta_0}\left(M^i(\theta_0)\right) < \infty,$$

where the integrability of $M^i(\theta_0)$ follows from Theorem 3.15 above. Hence,

$$\mathbb{E}_{\theta_0}\left(\omega_N(\eta)\right) \leq \frac{\eta}{N}\mathbb{E}_{\theta_0}\left(\sup_{\theta\in\Theta}\left\|\frac{d}{d\theta}\ell_N(\theta; \boldsymbol{X})\right\|\right) \leq K\frac{\eta}{N}N\mathbb{E}_{\theta_0}\left(M^i(\theta_0)\right) = K\eta,$$

which assures the weak consistency of $\hat{\theta}_N$. Now we will verify the asymptotic normality of $\hat{\theta}_N$, i.e., $\sqrt{N}\mathrm{vec}\left(\hat{\theta}_N - \theta_0\right) \longrightarrow \mathcal{N}\left(0, \mathcal{I}(\theta_0)^{-1}\right)$, with $\mathcal{I}(\theta_0)$ as the covariance matrix of the vectorized (one-sample) Score function. By the consistency of $\hat{\theta}_N$ and since the true value θ_0 is an inner point of Θ, $\mathbb{Q}_{\theta_0}\left(\hat{\theta}_N \in \mathrm{int}(\Theta)\right) \to 1$. Let $K_N(\theta) = -\frac{1}{N}\ell_N(\theta; \boldsymbol{X})$ and denote by $K_{N,\mu}$ the gradient of K_N w.r.t μ, similar for $K_{N,\Omega}$. The mean value theorem for vector-valued functions of several variables assures that

$$K_{N,\mu}(\hat{\theta}_N) - K_{N,\mu}(\theta_0) = \int_0^1 \mathcal{J}_{K_{N,\mu}}(\theta_0 + t(\hat{\theta}_N - \theta_0))dt\,(\hat{\theta}_N - \theta_0) =: J_{K_{N,\mu}}(\hat{\theta}_N, \theta_0)(\hat{\theta}_N - \theta_0)$$

with $\mathcal{J}_{K_{N,\mu}}(\theta) = \left(\frac{d}{d\mu}K_{N,\mu}(\theta)', \frac{d}{d\mathrm{vec}(\Omega)}K_{N,\mu}(\theta)'\right)$ being the Jacobian of $K_{N,\mu}(\theta)$. Analogously,

$$K_{N,\Omega}(\hat{\theta}_N) - K_{N,\Omega}(\theta_0) = \int_0^1 \mathcal{J}_{K_{N,\Omega}}(\theta_0 + t(\hat{\theta}_N - \theta_0))dt\,(\hat{\theta}_N - \theta_0) =: J_{K_{N,\Omega}}(\hat{\theta}_N, \theta_0)(\hat{\theta}_N - \theta_0),$$

with $\mathcal{J}_{K_{N,\Omega}}(\theta) = \left(\frac{d}{d\mu}K_{N,\Omega}(\theta)', \frac{d}{d\mathrm{vec}(\Omega)}K_{N,\Omega}(\theta)'\right)$. Let

$$\mathcal{I}_N(\theta, \theta_0) := \begin{bmatrix} J_{K_{N,\mu}}(\theta, \theta_0) \\ J_{K_{N,\Omega}}(\theta, \theta_0) \end{bmatrix} = \int_0^1 \mathcal{H}_{K_N}(\theta_0 + t(\theta - \theta_0))dt,$$

where $\mathcal{H}_{K_N}(\theta)$ is the Hessian of $K_N(\theta)$. Then

$$\frac{d}{d\theta}K_N(\hat{\theta}_N) - \frac{d}{d\theta}K_N(\theta_0) = \mathcal{I}_N(\hat{\theta}_N, \theta_0)(\hat{\theta}_N - \theta_0). \tag{3.13}$$

The law of large numbers (see also Theorem 3.16 and its proof) assures the convergence $\mathcal{H}_{K_N}(\theta_0) = \mathcal{I}_N(\theta_0, \theta_0) \xrightarrow{\mathbb{Q}_{\theta_0}} \mathcal{I}(\theta_0)$. In the following we prove that $\mathcal{I}_N(\hat{\theta}_N, \theta_0)$ converges to $\mathcal{I}(\theta_0)$ in $\mathbb{Q}_{\theta_0}$-probability, as $N \to \infty$, by showing that each entry of the difference matrix

$$\mathcal{I}_N(\hat{\theta}_N, \theta_0) - \mathcal{I}_N(\theta_0, \theta_0) = \int_0^1 \mathcal{H}_{K_N}(\theta_0 + t(\hat{\theta}_N - \theta_0))dt - \mathcal{H}_{K_N}(\theta_0)$$

converges to 0 in $\mathbb{Q}_{\theta_0}$-probability. This will only be verified for the first (upper left) entry

$$r_N := \int_0^1 \frac{d^2}{d\mu_1^2}K_N(\theta_0 + t(\hat{\theta}_N - \theta_0))dt - \frac{d^2}{d\mu_1^2}K_N(\theta_0),$$

the remaining ones can be treated analogously. We define for $\delta > 0$ the set

$$B_{\delta,N} = \{x \in C_T : \max_{0\le t\le 1}\left(\left\|\theta_0 + t(\hat{\theta}_N - \theta_0)\right\|\right) < \delta\}$$

and note that the consistency of $\hat{\theta}_N$ implies $\lim_{\delta\to 0}\lim_{N\to\infty}\mathbb{Q}_{\theta_0}(B_{\delta,N}) = 1$.

$$\begin{aligned}
\mathbb{E}_{\theta_0}\left(|r_N|\cdot \mathbb{1}_{B_{\delta,N}}\right) &\le \frac{1}{N}\mathbb{E}_{\theta_0}\left(\int_0^1 \left|\frac{d^2}{d\mu_1^2}\ell_N(\theta_0 + t(\hat{\theta}_N - \theta_0); \boldsymbol{X}) - \frac{d^2}{d\mu_1^2}\ell_N(\theta_0; \boldsymbol{X})dt\right|\cdot \mathbb{1}_{B_{\delta,N}}\right)\\
&\le \mathbb{E}_{\theta_0}\left(\int_0^1 \left|\frac{d^2}{d\mu_1^2}\ell(\theta_0 + t(\hat{\theta}_N - \theta_0); X^i) - \frac{d^2}{d\mu_1^2}\ell(\theta_0; X^i)dt\right|\cdot \mathbb{1}_{B_{\delta,N}}\right)\\
&\le \mathbb{E}_{\theta_0}\left(\sup_{0\le t\le 1}\left|\frac{d^2}{d\mu_1^2}\ell(\theta_0 + t(\hat{\theta}_N - \theta_0); X^i) - \frac{d^2}{d\mu_1^2}\ell(\theta_0; X^i)\right|\cdot \mathbb{1}_{B_{\delta,N}}\right)\\
&\le \mathbb{E}_{\theta_0}\left(f_{\theta_0}(\delta)\right),
\end{aligned}$$

with $f_{\theta_0}(\delta) = \sup_{\|\theta-\theta_0\|<\delta}\left|\frac{d^2}{d\mu_1^2}\ell(\theta; X^i) - \frac{d^2}{d\mu_1^2}\ell(\theta_0; X^i)\right|$, which is independent of N. The continuity of $\theta \mapsto \frac{d^2}{d\mu_1^2}\ell(\theta; X^i)$ assures that $f_{\theta_0}(\delta)$ converges to 0 as $\delta \to 0$. Moreover, $f_{\theta_0}(\delta)$ can be bounded uniformly, since $\frac{d^2}{d\mu d\mu'}\ell(\theta; X^i) = G^i(\Omega)$, $[\![G^i(\Omega)]\!] \le \frac{1}{M_{2,1}}$ and thus,

$$\sup_{\|\theta-\theta_0\|<\delta}\left|\frac{d^2}{d\mu_1^2}\ell(\theta; X^i)\right| \le K \sup_{\|\theta-\theta_0\|<\delta}[\![G^i(\Omega)]\!] \le K.$$

Hence, dominated convergence applies and we get

$$\lim_{\delta\to 0}\limsup_{N\to\infty}\mathbb{E}_{\theta_0}\left(|r_N|\cdot \mathbb{1}_{B_{\delta,N}}\right) = 0.$$

The convergence of $|r_N|$ to 0 in $\mathbb{Q}_{\theta_0}$-probability can now be concluded from $\lim_{\delta\to 0}\lim_{N\to\infty}\mathbb{Q}_{\theta_0}(B_{\delta,N}) = 1$. This completes the proof of the $\mathbb{Q}_{\theta_0}$-convergence of $\mathcal{I}_N(\hat{\theta}_N, \theta_0)$ to $\mathcal{I}(\theta_0)$. We introduce the sets

$$B_N^1 = \{x \in C_T : \hat{\theta}_N \in \mathrm{int}(\Theta)\}, \qquad B_N^2 = \{x \in C_T : \mathcal{I}_N(\theta_0, \theta_0)^{-1} \text{ exists }\}.$$

Note that - by definition of the MLE - we have $\frac{d}{d\theta}K_N(\hat{\theta}_N) = 0$ on B_N^1. The consistency of the MLE $\hat{\theta}_N$ assures $\lim_{N\to\infty} \mathbb{Q}_{\theta_0}(B_N^1) = 1$. As $\mathcal{I}_N(\hat{\theta}_N, \theta_0)$ converges to $\mathcal{I}(\theta_0)$, one can moreover conclude that $\lim_{N\to\infty} \mathbb{Q}_{\theta_0}(B_N^2) = 1$ and the continuous mapping theorem gives

$$\mathcal{I}_N(\hat{\theta}_N, \theta_0)^{-1} \mathbb{1}_{B_N^1 \cap B_N^2} \xrightarrow{\mathbb{Q}_{\theta_0}} \mathcal{I}(\theta_0)^{-1}.$$

From identity (3.13) and the fact that $\frac{d}{d\theta}K_N(\hat{\theta}_N) = 0$ on B_N^1 we conclude that

$$\mathcal{I}_N(\hat{\theta}_N, \theta_0)^{-1} \left(-\sqrt{N}\right) \frac{d}{d\theta} K_N(\theta_0) \cdot \mathbb{1}_{B_N^1 \cap B_N^2} = \sqrt{N}\left(\hat{\theta}_N - \theta_0\right) \cdot \mathbb{1}_{B_N^1 \cap B_N^2}.$$

Theorem 3.16 assures that $-\sqrt{N}\frac{d}{d\theta}K_N(\theta_0)$ converges in distribution to $\mathcal{N}(0, \mathcal{I}(\theta_0))$ and application of Slutsky's lemma finally gives

$$\sqrt{N}\left(\hat{\theta}_N - \theta_0\right) = -\mathcal{I}_N(\hat{\theta}_N, \theta_0)^{-1} \sqrt{N} \frac{d}{d\theta} K_N(\theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1}).$$

□

If $B(t,x)'\Gamma(t,x)^{-1}B(t,x) = b(t)$ is constant as a function of x, we have $V_i = \int_0^T b(t)dt =: V$. Previously, we derived that the MLEs $\hat{\mu}_N$ and $\hat{\Omega}_N$ are implicitly given by the system (3.9). We set $\overline{U}_N = \frac{1}{N}\sum_{i=1}^N U_i$ and we obtain the *explicit* expressions

$$\begin{aligned}
\hat{\mu}_N &= V^{-1}\overline{U}_N \\
\hat{\Omega}_N &= V^{-1}\left(\frac{1}{N}\sum_{i=1}^N (U_i - \overline{U}_N)(U_i - \overline{U}_N)' V^{-1} - I_d\right).
\end{aligned}$$

To study the asymptotic behavior of these estimators, we observe that

$$\begin{aligned}
U_i &= \int_0^T B(s, X_s^i)'\Gamma(s, X_s^i)^{-1}\left[dX_s^i - A(s, X_s^i)ds\right] \\
&= \int_0^T B(s, X_s^i)'\Gamma(s, X_s^i)^{-1}B(s, X_s^i)ds \cdot \phi^i + \int_0^T B(s, X_s^i)'\Gamma(s, X_s^i)^{-1}\Sigma(s, X_s^i)dW_s^i \\
&= V \cdot \phi^i + \underbrace{\int_0^T B(s, X_s^i)'\Gamma(s, X_s^i)^{-1}\Sigma(s, X_s^i)dW_s^i}_{=:M^i} = V \cdot \phi^i + M^i.
\end{aligned}$$

Using that $\mathbb{E}_\theta(U_i) = V\mu$, strong consistency of $\hat{\mu}_N$ follows immediately from the LLN. To study the consistency of $\hat{\Omega}_N$, we note that $\frac{1}{N}\sum_{i=1}^N (U_i - \overline{U}_N)(U_i - \overline{U}_N)'$ converges almost surely to $\mathbb{C}\mathrm{ov}_\theta(U_i)$ according to the LLN. We further observe that $\mathbb{C}\mathrm{ov}_\theta(M^i) = V$ and $\mathbb{C}\mathrm{ov}_\theta(\phi^i, M^i) = \mathbb{E}_\theta\left(\phi^i (M^i)'\right) = \mathbb{E}_\theta\left(\phi^i \mathbb{E}_\theta[(M^i)'|\phi^i]\right) = 0$, such that $\mathbb{C}\mathrm{ov}_\theta(U_i)$ is given by

$$\begin{aligned}
\mathbb{C}\mathrm{ov}_\theta(U_i) &= \mathbb{C}\mathrm{ov}_\theta(V\phi^i + M^i) \\
&= V\Omega V + \mathbb{C}\mathrm{ov}_\theta(M^i) + V\mathbb{C}\mathrm{ov}_\theta(\phi^i, M^i) + \mathbb{C}\mathrm{ov}(\psi^i, M^i)'V \\
&= V\Omega V + V.
\end{aligned}$$

Therefore, the MLE $\hat{\Omega}_N$ converges a.s. to $V^{-1}\left([V\Omega V + V]V^{-1} - I_d\right) = \Omega$ and is thus strongly consistent. As $\hat{\mu}_N$ is an average of the independent and identically distributed random vectors $V^{-1}U_i$ with mean vector μ, the standard multivariate central limit theorem assures that $\sqrt{N}\,(\hat{\mu}_N - \mu)$ is asymptotically centered Gaussian distributed. For asymptotic normality of (the vectorized version of) $\sqrt{N}\left(\hat{\Omega}_N - \Omega\right)$ it is enough to establish asymptotic normality of $\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}(U_i - \overline{U}_N)(U_i - \overline{U}_N)'\right)$, which follows from standard theory.

3.4.6 Auxiliary theorems

This section provides auxiliary results that are employed in proofs of this chapter. Most are well-known and stated without proof, as they can be found in standard literature, such as Gikhman and Skorokhod (1979), Karatzas and Shreve (1991), or Lipster and Shiryaev (2001).

Theorem A.1 (SDE solutions in case of random global Lipschitz constants)**.** *For $0 \leq t \leq T, \omega \in \Omega$, consider the r-dimensional SDE*

$$dX_t(\omega) = \alpha(t, X_t(\omega), \omega)dt + \Sigma(t, X_t(\omega))dW_t(\omega), \quad X_0(\omega) = Y(\omega), \tag{3.14}$$

with r-dimensional standard Brownian motion $(W, (\mathcal{F}_t)_{0\leq t\leq T})$, $Y \in L_2(\mathbb{P})$, and measurable functions $\alpha : [0,T] \times \mathbb{R}^r \times \Omega \to \mathbb{R}^r$ and $\Sigma : [0,T] \times \mathbb{R}^r \to \mathbb{R}^{r\times r}$. Assume that for all $x \in \mathbb{R}^r$ the process $(t,\omega) \mapsto \alpha(t,x,\omega)$ is adapted to $(\mathcal{F}_t)_{0\leq t\leq T}$. Assume moreover that $\int_0^T \|\alpha(t,x,\omega)\|^2 + [\![\Sigma(t,x)]\!]^2 dt < \infty$ for $\mathbb{P}$-a.a. ω. Let K_1, K_2 be finite $\mathcal{F}_0$-measurable random variables and suppose that for $\mathbb{P}$-a.a. $\omega \in \Omega$ the following holds:

$$\|\alpha(t,x,\omega)\| + [\![\Sigma(t,x)]\!] \leq K_1(\omega)\,(1 + \|x\|)$$
$$\|\alpha(t,x,\omega) - \alpha(t,z,\omega)\| + [\![\Sigma(t,x) - \Sigma(t,z)]\!] \leq K_2(\omega)\,\|x - z\|\,.$$

Then there exists a unique (a.s.) continuous, adapted process solution $X = (X_t)_{0\leq t\leq T}$ to the SDE above.

Proof. We start with the existence proof. Let $K = K_1 + K_2$. For $M \in \mathbb{N}$ introduce the sets $\Omega_M = \{\omega \in \Omega : K(\omega) \leq M\}$. Then $\Omega_M \subset \Omega_{M+1} \subset \ldots$ and $\mathbb{P}(\Omega_M) \to 1$, since $K < \infty$ a.s. Let us define the truncated function $\alpha_M := \alpha \mathbb{1}_{\Omega_M}$. Then $\alpha_M(\cdot, x, \cdot)$ is adapted (since K is $\mathcal{F}_0$-measurable) for every $x \in \mathbb{R}^r$ and it satisfies the standard non-random Lipschitz conditions with Lipschitz constant M. Moreover, $F_M = F_{M+1} = \ldots = F$ on Ω_M. Standard existence and

uniqueness results assure the existence of a unique continuous, adapted process $X^{(M)}$ defined on $[0,T]$ and satisfying for almost all $\omega \in \Omega$

$$X_t^{(M)}(\omega) = Y(\omega) + \int_0^t \alpha_M(s, X_s^{(M)}(\omega), \omega)ds + \int_0^t \Sigma_M(s, X_s^{(M)}(\omega)))dW_s(\omega).$$

Due to the uniqueness of solutions, we have $X^{(M)} = X^{(M+1)} = X^{(M+2)} = \dots$ on Ω_M. Since $\mathbb{P}(K < \infty) = 1$, there is $\Omega_0 \subset \Omega, \mathbb{P}(\Omega_0) = 1$, such that for all $\omega \in \Omega_0$ there is an integer $M_0(\omega)$ such that for all $M \geq M_0(\omega)$ one has $K(\omega) \leq M$. We define

$$X(\omega) := X^{(M_0(\omega))}(\omega)\mathbb{1}_{\Omega_0}(\omega).$$

Then for any $M \in \mathbb{N}$ we have $X = X^{(M)}$ on Ω_M and since $\alpha_M = \alpha$ on Ω_M, the process X satisfies (3.14) on Ω_M. Now let $M \to \infty$ and use that $\mathbb{P}(\Omega_M) \to 1$ to conclude that (3.14) holds a.s. on Ω. Since for almost all $\omega \in \Omega$ there is an $M(\omega)$ such that $X(\omega) = X^{(M(\omega))}(\omega)$, continuity and adaptedness of X follows from the corresponding properties of $X^{(M)}$. To verify uniqueness, assume that $X, \tilde{X}$ are continuous, adapted solutions to (3.14). Then on Ω_M we have $X = \tilde{X} = X^{(M)}$ (due to uniqueness of solutions). Since $\mathbb{P}(\Omega_M) \to 1$ we conclude that $X = \tilde{X}$ a.s.

□

The following result is a Corollary to Theorem 2.4.1 in Mao (2007).

Theorem A.2 (Existence of moments of the unique solution)**.** *Assume the setting of Theorem A.1 and let X be the unique, continuous and adapted solution to* (3.14)*. Let $p \geq 2$ and suppose that the initial condition Y in* (3.14) *satisfies $Y \in L^p(\Omega)$. If K_1 is deterministic, X satisfies*

$$\mathbb{E}\left(\|X_t\|^p\right) \leq 2^{\frac{p-1}{2}}\left(1 + \mathbb{E}(\|Y\|^p)\right) e^{p\left[\sqrt{K_1} + K_1(p-1)/2\right]t}.$$

Lemma A.3. *Let $\mathbb{Q}_\theta, \mathbb{Q}_0$ be two probability measures on $(C_T, \mathcal{C}_T)$ and assume that $\mathbb{Q}_\theta \ll \mathbb{Q}_0$. Let $U : C_T \to S$ be a measurable mapping into some measure space $(S, \mathfrak{S})$ and let $X : C_T \to C_T$ be a random variable whose distribution is $\mathbb{Q}_\theta$. Then the distribution $\mathbb{Q}_\theta^{U(X)}$ of $U(X)$ has a density w.r.t. $\mathbb{Q}_0^{U(X)}$ and this density is given by*

$$\frac{d\mathbb{Q}_\theta^{U(X)}}{d\mathbb{Q}_0^{U(X)}}(u) = \mathbb{E}_{\mathbb{Q}_0}\left[\frac{d\mathbb{Q}_\theta}{d\mathbb{Q}_0}(X)|U(X) = u\right].$$

Theorem A.4. *Let X be a random variable and $M_X(t) = \mathbb{E}\left(e^{tX}\right) \in \overline{\mathbb{R}}_+$ its moment generating function. Assume there exist $t_0 < 0 < t_1$ such that $M_X(t_0) < \infty$ and $M_X(t_1) < \infty$. Then the moments of all orders of X exist and are finite.*

Theorem A.5 (Kolmogorov's continuity criterion, Revuz-Yor Th.2.1)**.** *Let $X = (X_t)_{t\in[0,1]^d}$ be a Banach-valued process for which there exist constants $\gamma, \kappa, K > 0$ such that*

$$\mathbb{E}\left(\|X_t - X_s\|^\gamma\right) \leq K|t-s|^{d+\kappa}.$$

Then there exists a (everywhere) continuous modification $\tilde{X}$ of X, i.e. $\tilde{X}(\omega)$ is a continuous function on $[0,T]$ for each $\omega \in \Omega$. Even more, $\tilde{X}$ is Hölder continuous of order α for all $\alpha \in [0, \kappa/\gamma)$.

Theorem A.6 (Burkholder-Davis-Gundy)**.** *Let $(M_t)_{0\leq t\leq T}$ be a continuous real-valued local martingale with $M_0 = 0$. For every $0 < p < \infty$ there are constants k_p, K_p such that*

$$k_p\mathbb{E}\left(\langle M\rangle_T^{p/2}\right) \leq \mathbb{E}\left(\sup_{0\leq t\leq T}|M_t|^p\right) \leq K_p\mathbb{E}\left(\langle M\rangle_T^{p/2}\right).$$

Theorem A.7 (Multidimensional Burkholder-Davis-Gundy)**.** *Let W be a m-dimensional Wiener process, $H = (H^{ij})$ a matrix-valued $(d \times m)$-dimensional stochastic process s.t. each component is adapted to the filtration generated by W. Denote by H^i the i-th row of H and define $M_t^i = \int_0^t H_s^i dW_s$ for $i = 1, \ldots, d$. Then $\langle M^i\rangle_t = \int_0^t \left\|H_s^i\right\|^2 ds$ is the quadratic variation and $M = (M^1, \ldots, M^d)'$ is a d-dimensional martingale. Let $p \geq 2$. There are constants $c_p, \tilde{c}_p$ such that*

$$c_p\mathbb{E}\left(\left[\sum_{i=1}^d \left|M_t^i\right|^2\right]^{p/2}\right) \leq \mathbb{E}\left(\left[\sum_{i=1}^d \langle M^i\rangle_t\right]^{p/2}\right) \leq \tilde{c}_p\mathbb{E}\left(\left[\sum_{i=1}^d \left|M_t^i\right|^2\right]^{p/2}\right).$$

Theorem A.8 (Differentiability of parameter integrals (Amann and Escher, 2009))**.** *Let $E = (E, \|\cdot\|)$ be a Banach space and $(X, \mathcal{A}, \mu)$ a complete σ-finite measure space. Let $U \subseteq \mathbb{R}^n$ be open and $f : X \times U \to E$ such that*

- *$x \mapsto f(x, \varphi)$ is integrable for each φ*
- *$\varphi \mapsto f(x, \varphi)$ is continuously differentiable for μ-a.a. $x \in X$*

- $\left\| \frac{d}{d\varphi_j} f(x,\varphi) \right\| \leq g(x)$ *for all* (x,φ) *and all* $j = 1, \ldots, d$

Then $\varphi \mapsto \int_X f(x,\varphi) d\mu(x)$ *is continuously differentiable and one can differentiate under the integral sign.*

Theorem A.9. *Let* $f : [0,T] \times \mathbb{R}^d \to \mathbb{R}$ *be continuous in both arguments. Then for any* $R > 0$ *the function* $F : [-R,R]^d \to \mathbb{R}$ *defined by* $F(\varphi) = \int_0^T f(s,\varphi) ds$ *is continuous.*

Proof. Let $\epsilon > 0$ be arbitrary. Since f is continuous on $[0,T] \times [-R,R]^d$, it is uniformly continuous, i.e. for all $\tilde{\epsilon} > 0$ there is a $\delta > 0$ such that whenever $\|(s,\varphi) - (\tilde{s},\tilde{\varphi})\| < \delta$ we have that $|f(s,\varphi) - f(\tilde{s},\tilde{\varphi})| < \tilde{\epsilon}$. Let $\tilde{\epsilon} = \frac{\epsilon}{T}$. By uniform continuity there is a $\delta > 0$ such that $|f(s,\varphi) - f(s,\varphi+h)| < \tilde{\epsilon} = \frac{\epsilon}{T}$ whenever $\|h\| < \delta$. Hence, for h such that $\|h\| < \delta$ we get

$$|F(\varphi+h) - F(\varphi)| \leq \int_0^T |f(s,\varphi) - f(s,\varphi+h)| \, ds < T\frac{\epsilon}{T} = \epsilon.$$

□

Papers and Manuscripts

I - Automated analysis of song structure in complex birdsongs

Published in *Animal Behaviour*, 112 (2016)
DOI: 10.1016/j.anbehav.2015.11.013

Mareile Große Ruse
Department of Mathematical Sciences
University of Copenhagen, Copenhagen, Denmark

Dennis Hasselquist
Department of Biology
Lund University, Lund, Sweden

Bengt Hansson
Department of Biology
Lund University, Lund, Sweden

Maja Tarka
Center for Biodiversity Dynamics (CBD)
Norwegian University of Science and Technology (NTNU), Trondheim, Norway

Maria Sandsten
Centre for Mathematical Sciences
Lund University, Lund, Sweden

Keywords: ambiguity spectrum, automated song recognition, birdsong, clustering, great reed warbler, multitaper, song analysis, syllable detection

Animal Behaviour 112 (2016) 39–51

Contents lists available at ScienceDirect

Animal Behaviour

journal homepage: www.elsevier.com/locate/anbehav

Commentary

Automated analysis of song structure in complex birdsongs

Mareile Große Ruse [a, *], Dennis Hasselquist [b], Bengt Hansson [b], Maja Tarka [c], Maria Sandsten [d]

[a] Department of Mathematical Sciences, University of Copenhagen, Copenhagen, Denmark
[b] Department of Biology, Lund University, Lund, Sweden
[c] Center for Biodiversity Dynamics (CBD), Norwegian University of Science and Technology (NTNU), Trondheim, Norway
[d] Centre for Mathematical Sciences, Lund University, Lund, Sweden

ARTICLE INFO

Article history:
Received 23 June 2015
Initial acceptance 20 July 2015
Final acceptance 13 October 2015
Available online
MS. number: 15-00536R

Keywords:
ambiguity spectrum
automated song recognition
birdsong
clustering
great reed warbler
multitaper
song analysis
syllable detection

Understanding communication and signalling has long been strived for in studies of animal behaviour. Many songbirds have a variable and complex song, closely connected to territory defence and reproductive success. However, the quantification of such variable song is challenging. In this paper, we present a novel, automated method for detection and classification of syllables in birdsong. The method provides a tool for pairwise comparison of syllables with the aim of grouping them in terms of their similarity. This allows analyses such as (1) determining repertoire size within an individual, (2) comparing song similarity between individuals within as well as between populations of a species and (3) comparing songs of different species (e.g. for species recognition). Our method is based on a particular feature representation of song units (syllables) which ensures invariance to shifts in time, frequency and amplitude. Using a single song from a great reed warbler, *Acrocephalus arundinaceus*, recorded in the wild, the proposed algorithm is evaluated by means of comparison to manual auditory and visual (spectrogram) song investigation by a human expert and to standard song analysis methods. Our birdsong analysis approach conforms well to manual classification and, moreover, outperforms the hitherto widely used methods based on mel-frequency cepstral coefficients and spectrogram cross-correlation. Thus, our algorithm is a methodological step forward for analyses of song (syllable) repertoires of birds singing with high complexity.

Birdsong is among the most prominent and widespread avian behaviours, dominating the audial environment in spring and early summer. Considerable research effort has been devoted to questions such as how birds sing and why birds have such elaborate songs (Catchpole & Slater, 1995; Miller & Kroodsma, 1996). Central to most research questions involving birdsong is the need for classifying, comparing and quantifying sounds in the song within and among individuals. Typical questions include how song repertoire size or song variability influence mate choice and male–male competition (e.g. Hasselquist, Bensch, & von Schantz, 1996; Horn & Falls, 1996; Searcy & Yasukawa, 1996). This involves comparisons of songs within an individual to estimate song complexity (Catchpole, 1976), comparing song similarity among individuals within a species (e.g. neighbour song matching, Falls, 1985; Horn & Falls, 1988) and variation between individuals within a population (Slater, Clements, & Goodfellow, 1984) and over time (Lehtonen, 1983), investigating geographical variation (Catchpole & Rowell, 1993) and song dialects (Espmark, Lampe, & Bjerke, 1989; McGregor, 1980; Mundinger, 1980), as well as variation between species, allowing species recognition (Kreutzer & Güttinger, 1991; Martens, 1996; Miller, 1996). Studies assessing vocal development and song learning endeavour to quantify similarities (imitation) between the song of a young bird and its tutor (Kroodsma & Konishi, 1991; Nottebohm, 1991; Slater & Ince, 1982). Classification of song sounds and estimation of song complexity (i.e. song repertoire size) can be conducted either (1) on whole song strophes in species with low to medium song complexity such as the chaffinch, *Fringilla coelebs* (Slater, 1983) and the American redstart, *Setophaga ruticilla* (Lemon, Cotter, MacNally, & Monette, 1985) or (2) on smaller sound entities, such as syllables, which are discrete sound units that build up (often a large number of different) song strophes in species with higher song complexity such as the great reed warbler (GRW), *Acrocephalus arundinaceus* (Hasselquist, 1998; Hasselquist et al., 1996; Węgrzyn & Leniowski, 2010) and the pied

* Correspondence: M. Große Ruse, Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, Copenhagen 2100, Denmark.
E-mail address: mareile@math.ku.dk (M. Große Ruse).

http://dx.doi.org/10.1016/j.anbehav.2015.11.013
0003-3472/

flycatcher, *Ficedula hypoleuca* (Eriksen, Slagsvold, & Lampe, 2011; Lampe & Espmark, 2003).

The hitherto standard methods to classify song entities (syllables) has been by means of the audial and visual comparison of syllables (see e.g. Catchpole, 1976; Hasselquist et al., 1996; Węgrzyn, Leniowski, & Osiejuk, 2010), where the latter is often conducted based on syllable spectrograms (Adret, Meliza, & Margoliash, 2012; Węgrzyn & Leniowski, 2010). Unfortunately, these approaches are often time consuming, prone to observer bias and subjectivity, non-numerical (making statistical analyses problematic) and perform less well on songs with large syllable repertoires or with complex structures of song strophes/syllables (Clark, Marler, & Beeman, 1987; Williams, 1993; Williams & Slater, 1991). The algorithm we propose enables an automated objective classification of birdsongs and thereby facilitates the assessment of the song repertoire of an individual bird and its temporal development, as well as comparisons of song structures among birds within and between different populations.

Automated song analysis is typically conducted by subdividing a song into smaller entities (e.g. syllables or even larger song sections as in Tchernichovski, Nottebohm, Ho, Pesaran, & Mitra, 2000) and representing them in terms of features, i.e. a collection of characterizing properties. These features can be intuitive parameters such as the time length of a unit, its power or pitch or more involved quantities, such as mel-frequency cepstral coefficients (MFCC) or wavelet coefficients (see below). The selection of features, however, should be guided by the purpose of the study, properties of the data and not least by available computational resources. The feature representation summarizes important characteristics of the song units, which facilitates a comparative analysis in terms of a similarity measure that operates on the feature basis. If the differences of interest between syllables are known in advance and are sufficiently pronounced, the representing features should be chosen such that they are able to reflect these differences prominently. If, for example, one factor of interest is a syllable's length, the time duration will be the natural choice and should be included in the feature representation. However, if there is no prior knowledge of the underlying factors for classification or if syllable characteristics are very complex and not straightforward to capture, the purported feature space should be sufficiently rich in order to facilitate detection of various syllable characteristics. Methods for song analysis that have been proposed in the literature differ mainly in the way song units are chosen, which features are extracted and how similarity between features of song units is assessed. The choice of small entities such as single syllables as the basic building blocks facilitates song complexity analysis of songs with highly variable strophes, such as those of most European and Asian warblers, flycatchers, thrushes, wrens and chats. Moreover, sections of a recording that are affected by substantial background noise can easily be discarded, which proves beneficial if only field recordings are available. If, however, the whole recording is confounded by noise, it may be hard to tell subsequent syllables apart, which hampers an unambiguous definition/detection of syllables.

The most well-known technique in the context of birdsong analysis is the cross-correlation approach applied to the spectrogram (SPCC; Keen, Ross, Griffiths, Lanzone, & Farnsworth, 2014). This method, however, suffers from sensitivity to natural jitters of components at time and frequency locations as well as sensitivity to noise. Tchernichovski, Lints, Mitra, and Nottebohm (1999) therefore applied the more robust multitaper (MT) spectrograms for syllable-based feature extraction. Kogan and Margoliash (1998) showed that spectrogram-based features are inferior to a syllable representation by means of MFCC when recordings are substantially affected by background noise. Syllable representation by time-varying sinusoids was applied to recognition of bird species with relatively simplistic songs by Härmä (2003), while Somervuo and Härmä (2004) combined this approach with a clustering of syllables in terms of the *k*-means algorithm. In a comparative study, Somervuo, Härmä, and Fagerlund (2006) found that the sinusoidal model approach was clearly outperformed by the more complex MFCC-based syllable representation. A combination of the latter and some descriptive parameters for syllable representation was investigated by Fagerlund (2007) and by Trifa, Kirschel, Taylor, and Vallejo (2008) in the context of species recognition. For more complex birdsong syllables the comparably simple models such as the sinusoidal approach and the MFCC representation may fail to capture central information of the signals and more sophisticated representations such as wavelet decompositions have been considered (Selin, Turunen, & Tanttu, 2007). (For a recent comparison of various methods, such as SPCC, dynamic time warping and pitch-frequency analysis, see Keen et al., 2014; Meliza, Keen, & Rubenstein, 2013.) Algorithms employing some of the previously mentioned techniques have been implemented and made available as ready-to-use programs, among them Sound Analysis Pro (Tchernichovski & Mitra, 2004), Luscinia (Lachlan, 2007), Avisoft-SASLab Pro (Specht, 2004) or Praat (Boersma & Weenink, 2001).

For the within-species analysis of songs with elaborate complexity (e.g. such as those of the GRW), however, techniques with an additional level of sensitivity are required.

In this paper, we propose a fully automated method tailored to syllable-based, within-species analysis of field recordings of complex birdsongs. The algorithm, which allows an unbiased, reproducible and reliable song analysis, is a three-step procedure, comprising syllable detection, representation and comparison (with further clustering, if required). The novelties in our approach are the usage of the ambiguity spectrum (a transformation of the spectrogram and also called the Doppler-lag spectrum) for feature extraction and a novel similarity measure for subsequent syllable comparison.

The advantage of the ambiguity spectrum as opposed to the popular standard spectrogram is its invariance to time and frequency shifts of syllables, as the ambiguity spectrum is always centred at zero Doppler frequency and time lag (Boashash, 2003). It therefore focuses solely on the relations between different time and frequency components in a syllable. As already successfully employed by, for example, Meliza et al. (2013) and Tchernichovski et al. (1999), as well as in our preliminary study (Sandsten, Tarka, Caissy-Martineau, Hansson, & Hasselquist, 2011), we used MTs for noise-robust spectrogram estimation. Syllable clustering is achieved by means of a hierarchical clustering algorithm where the number of clusters is objectively estimated by the Silhouette quality criterion. The proposed method, as well as comparative established approaches based on MFCC and SPCC, is evaluated on a real data set by comparison to a 'ground truth' given by a human expert (D.H.) with long experience in GRW song syllable analysis (Hasselquist, 1998; Hasselquist et al., 1996). All computations were conducted in MATLAB (MATLAB and Statistics Toolbox Release 2012b, The MathWorks, Inc., Natick, MA, U.S.A.) and we can provide a ready-to-use code upon request.

METHODS

Evaluation of the algorithm by comparison to human expert clustering and to methods based on MFCC and SPCC is conducted by means of three examples of increasing complexity. The underlying data represent the type of data typically obtained in the challenging setting of field recordings of complex birdsongs. In a first example, we use our algorithm and two alternative approaches (MFCC and SPCC) to recover a classification of syllables into two visually well-separated groups. The second example generalizes this setting to a

more realistic scenario by dropping the assumption that the syllable set is composed of exactly two classes. Instead, a clustering algorithm is applied which estimates the number of classes from the data and which will then group the syllables accordingly. The third example extends the previous clustering problem to the highly nontrivial and biologically very interesting task of clustering the syllables of a whole sequence of song strophes of a male GRW (236 s long and comprising 433 syllables).

Key Terms

Song strophe

A recording (referred to as a song in this text) is usually 3–10 min long and includes typically 25–40 song strophes, each of which is composed of approximately 10–20 smaller sound units, the syllables. Subsequent strophes are separated by a period of silence/no singing in which only noise is perceptible.

Song syllable

Syllables are more or less continuous sound sections separated by short silent periods and are the building blocks of a song strophe. A syllable in a GRW song has a duration of about 50–300 ms. Within a song strophe of the GRW, a syllable of certain type is usually repeated 1–10 times in a row. Syllables are the units of interest in the proposed song analysis tool and their clustering/classification is conducted by means of pairwise syllable comparisons.

Double syllable

A double syllable is a syllable containing two (usually repeated) or three parts (kack-a-kack) and is a common phenomenon in the song structure of GRWs.

Feature vector

To facilitate clustering analysis of a set of syllables, each syllable is represented by a feature vector, i.e. a collection of characterizing numbers (features). The quality of a feature vector is determined by its ability to discriminate between syllables from different groups.

Song Recording and Ethical Note

Each year within the period 1987–2010, songs were recorded of almost all GRW males that held a territory (for more than 7 days) in Lake Kvismaren (Hasselquist, 1998; Hasselquist et al., 1996). Recordings were done using a Telinga parabola with an attached Telinga Stereo DAT microphone used in mono mode (Telinga Microphones, Sweden) and a SONY cassette tape recorder (SONY TC-D5M). Sounds were later digitalized using a Tascam 322 cassette deck and a Lynx Aurora 16 soundcard. The sampling rate was 44.1 kHz and the bit depth 16 bits. The field recordings were made at a distance of about 10–60 m on males singing intensive (mate attraction) long songs (Catchpole, 1983; Hasselquist & Bensch, 1991). We avoided windy and rainy conditions and allowed males to recover a high song intensity after the initial approach of the human observer.

We were always cautious to approach a singing male slowly and carefully, to avoid having the male stop singing or fly off. In most cases, a singing male resumed intensive long song within a few minutes after our approach. Recordings were usually made from a small boat or canoe, which further reduced the disturbance when approaching the singing bird.

Syllable Detection

Syllable detection is conducted by means of two power-smoothing filters (moving averages). A longer (default 360 ms) filter $P_{\text{long}}(t)$ is responsible for noise reduction and determines a time-varying threshold, while a shorter (default 90 ms) filter $P_{\text{short}}(t)$ is used for detection of those samples that build up to a syllable. The syllable detection decision at each sample (time value) is based on whether

$$P_{\text{short}}(t) > P_{\text{long}}(t) + \left(1 - \frac{l_{\text{sens}}}{100}\right) \max P_{\text{long}}. \quad (1)$$

Here $\max P_{\text{long}}$ is the maximum value of $P_{\text{long}}(t)$ when t varies between 1 and L, where L is the strophe length (number of samples) and $l$sens is the sensitivity of the detector as a percentage (default 99), constraining the power of the short filter to be somewhat above the power of the long filter. With $l_{\text{sens}} = 100$, all small changes in the level of the recorded signal at the beginning and end of the strophe will be erroneously detected as syllables of weak power and short duration. When using a sensitivity of 99%, all syllables in a strophe with small disturbances, e.g. Fig. 1a, will be detected. Further lowering the sensitivity to, for example, $l_{\text{sens}} = 90$, creates an even less sensitive detector, where the low-level syllables especially in the beginning of the strophe as well as low-level disturbances will be discarded, for example that seen between syllables 7 and 8 in Fig. 1b. For recordings including a lot of noise, such as wind disturbance and songs from other individuals, a lower detection level is recommended, although this may lead to exclusion of less strongly pronounced syllables. If the time distance between consecutive, detected samples is smaller than a minimum allowed distance between syllables (default 60 ms), the detected samples are assumed to belong to the same syllable. The start and end time points of a detected syllable are extended backwards and forwards to include the weaker start and end of the signal (default ±60 ms). Choosing the extension to be less than or equal to the minimum allowed distance between syllables will ensure that no parts of neighbouring syllables will be included.

Syllable Representation

The basis of the syllable representation is a frequency and time shift-invariant transformation of the MT spectrogram, the filtered ambiguity spectrum. The spectrogram based on MTs is an improvement of the traditional spectrogram in terms of robustness as the variance of the resulting estimate of a syllable's time-frequency image is typically lower. Here we first recall the definition of the MT spectrogram and then explain the filtered ambiguity spectrum and describe how the filtered ambiguity spectrum of each syllable is used for feature extraction and how we assess the similarity of two syllables, based on their respective feature representation.

Multitaper spectrogram

Birdsong syllables are typically visualized by means of a spectrogram, based on a single (Hanning) window. Such a representation is depicted in Fig. 2a–d, in which two syllables are shown along with their corresponding Hanning window spectrograms (the colours are in logarithmic (dB) scale). This windowed spectrogram has a good time and frequency resolution, but suffers from noise sensitivity. The Welch spectrogram (Welch, 1967) achieves increased robustness by averaging spectra of partly overlapping data sequences, and an overlap of 50% has been shown to be appropriate from resolution and variance aspects. A further improvement in terms of leakage, resolution in frequency and

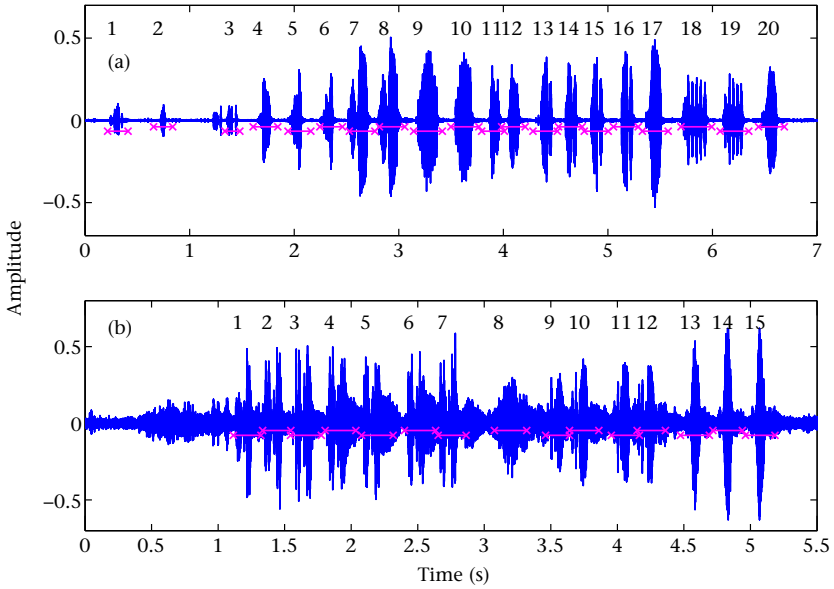


Figure 1. Strophes with detected syllables (marked with purple lines bordered with crosses and numbered). (a) Song strophe with small disturbances, $l_{sens} = 99\%$; (b) song strophe with large disturbances, $l_{sens} = 90\%$.

variance (Bronez, 1992) is achieved by using multiple windows, so-called multitapers (MTs), as introduced by Thomson (1982). In this approach all tapers make use of the information from the whole data set, and, provided certain properties are satisfied, the averaging of M different windowed spectrograms reduces the variance of the resulting estimate by up to a factor M, compared to a single-window spectrogram. Hermite function multitapers (Daubechies, 1988) give an additional improvement of the Thomson MTs in terms of resolution in time and frequency (Bayram & Baraniuk, 1996; Hansson-Sandsten, 2011; Xu, Haykin, & Racine, 1999), and is the method of choice for our proposed algorithm.

Examples of the Hermite function MT spectrograms using $M = 8$ multitapers are shown in Fig. 2e, f. When comparing these to the corresponding standard spectrograms (single Hanning window) shown in Fig. 2c, d, one can see that multitapering leads to a lower resolution but gives a clearer view of the signal components with better noise suppression (lower variance). The Hanning window has by definition a certain length and a frequency bandwidth that is sometimes measured as the frequency main lobe of the window (which is known to be $4/T$ for the Hanning window, where T is the actual length of the window). For the Hermite function MTs, it is not possible to define the time width in the same way, as the Hermite functions are of infinite length and the main lobe width does not necessarily relate to the Hanning main lobe. Instead, we define the time width of a single Hermite window as the time interval in which 99% of the energy is located. A corresponding definition of frequency width is to use the frequency interval that contains 99% of the window spectrum energy. As each of the M Hermite windows has different time and frequency widths, the time and frequency width of an M window Hermite MT spectrogram is then defined as the corresponding width of the Mth Hermite window as this window has the largest values. With a larger value of M, that is with more tapers, the time and frequency resolution of the corresponding final estimate will decrease. The time and frequency widths for the results shown in Fig. 2e, f are 13.4 ms and 883 Hz. The corresponding time width of the Hanning window in Fig. 2c, d is 4.17 ms and the frequency width is 474 Hz.

Filtered ambiguity spectrum

Even though the MT spectrogram is more robust to noise than the standard spectrogram and thus smooths out small jitters that could lead to misclassification of syllables, SPCC based on these improved spectrograms may still fail. For illustration of a problematic setting in this context, two syllables and their MT spectrograms are depicted in Fig. 3a–d. It can be seen that the general syllable structures are highly similar, suggesting that both syllables are realizations of the same syllable type (and thus should be classified as being equal). However, their spectrograms reveal an unequal number of strong components in the syllables (five in syllable C and six in syllable D). Thus, correlating the corresponding two single Hanning window spectrograms (Keen et al., 2014) will give a result of 0.67, which is far from being close to one (exact similarity). The reason is that, because of the different number of strong components, there are several possible positions in which the correlation is fairly high but none in which the two images match totally.

To circumvent those problematic phenomena, we instead concentrate on the filtered ambiguity spectrum (exemplified in Fig. 3e, f), which arises from the spectrogram by application of the Fourier transform (FT) in the two different directions. The transformation causes an invariance with respect to time and frequency shifts (Boashash, 2003) such that only time and frequency differences between syllable components are visible; the component's exact position in the time-frequency plane and the number of components are disregarded. This explains the close resemblance of the ambiguity spectra depicted in Fig. 3e, f.

Feature extraction

A song's syllable repertoire can be assessed by arranging syllables in clusters, where similar syllables are grouped together and distinct ones are divided into distinct clusters. A more effective and robust clustering can be facilitated by representing syllables by features. A first step to enhanced robustness was taken by application of MT techniques to spectrogram estimation. Transformation to the ambiguity domain eliminates the influence of time or

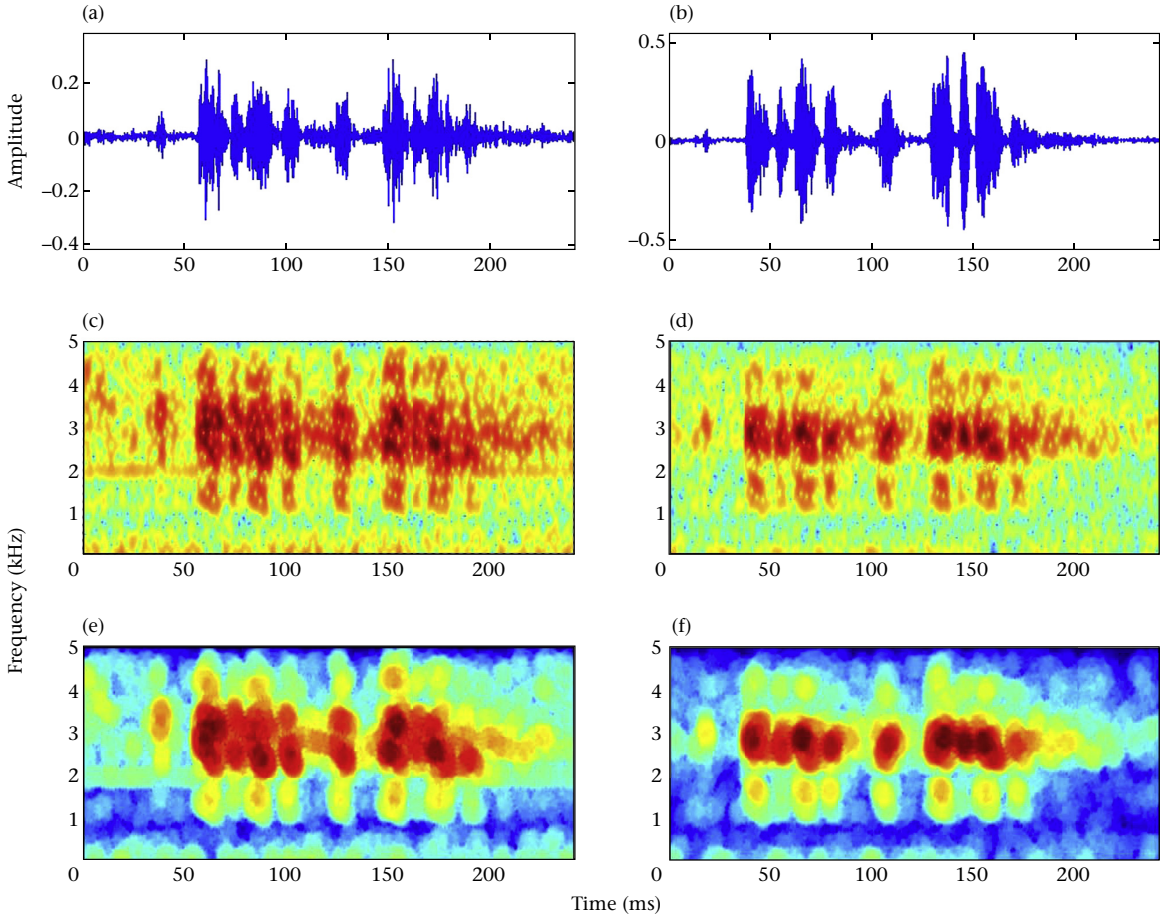


Figure 2. (a, b) Example of two similar syllables from one song. (a) Syllable A; (b) syllable B. (c, d) The corresponding spectrograms (dB scale) with time and frequency width 4.17 ms and 474 Hz for (c) syllable A and (d) syllable B. (e, f) The multitaper spectrograms (dB scale) with time and frequency width 13.4 ms and 883 Hz and $M = 8$ windows for (e) syllable A and (f) syllable B.

frequency shifts. To further reduce the influence of small jitters, such as the small frequency differences in Fig. 3c, d, and to compress information, we apply the singular value decomposition (SVD) to the ambiguity matrix. A syllable will be represented by only the first pair of singular vectors (a left and a right singular vector) of its ambiguity spectrum. Note that the syllable's amplitude information is found in the singular value, which will be disregarded. In that way, the information is normalized and the original syllable amplitudes will not influence the result.

The similarity between two syllables, both represented by their first left and right ambiguity singular vectors, is assessed by means of a similarity measure. Commonly used similarity measures are the cosine similarity, Pearson correlation or the extended Jaccard measure (Maimon & Rokach, 2005). Distance measures can also be used to define similarity measures and among the distances commonly used in clustering context are the Euclidean distance, L_1 (or Manhattan) distance, maximum distance, the binary distance or the Canberra distance. The similarity measure we propose in this work is related to the cosine similarity (for a definition of cosine similarity see e.g. Zaki & Meira, 2014).

For (column) vectors x, y, $\langle x, y\rangle = x^T y$ denotes their inner product and the superscript T indicates the transpose of a vector. For a syllable s we write $\widehat{u}$, $\widehat{v}$ for first left and right singular vectors obtained from the SVD of the estimated filtered ambiguity matrix. The similarity of two syllables $s^{(i)}$, $s^{(j)}$ in frequency and time structure is calculated as $\beta(s^{(i)}, s^{(j)}) = \min(\langle \widehat{u}^{(i)}, \widehat{u}^{(j)}\rangle, \langle \widehat{v}^{(i)}, \widehat{v}^{(j)}\rangle)$ and syllables are considered as similar if $\beta(s^{(i)}, s^{(j)}) \approx 1$, while nonsimilarity is inferred from $\beta(s^{(i)}, s^{(j)}) \approx 0$. The intuition behind this definition is the following. Singular vectors always have length one, i.e. $\langle \widehat{u}, \widehat{u}\rangle = \langle \widehat{v}, \widehat{v}\rangle = 1$. Now if a vector $\widetilde{u}$ is very similar to $\widehat{u}$, then one would expect that $\langle \widetilde{u}, \widehat{u}\rangle \approx 1$. On the other hand, if $\widetilde{u}$ is very different, say even perpendicular to $\widehat{u}$, then their inner product would approximately be zero, $\langle \widetilde{u}, \widehat{u}\rangle \approx 0$. Taking the minimum of both quantities in the definition of the similarity measure corresponds to considering the worst case scenario: Two syllables are only declared similar if they are similar in both time and frequency direction. Deviation in one direction yields nonsimilarity.

An example is seen in Fig. 4, in which the left and right singular vectors are depicted for the two syllables C and D of Fig. 3, with blue and red colours, respectively. It is seen that not only are the lag shapes very similar (close to one) but also the basic Doppler shapes resemble each other closely. The similarity measure in this case is $\beta(s^{(C)}, s^{(D)}) \approx 0.96$.

Syllable Clustering

Partitioning objects into different groups is usually referred to as classification or clustering, depending on whether the groups are

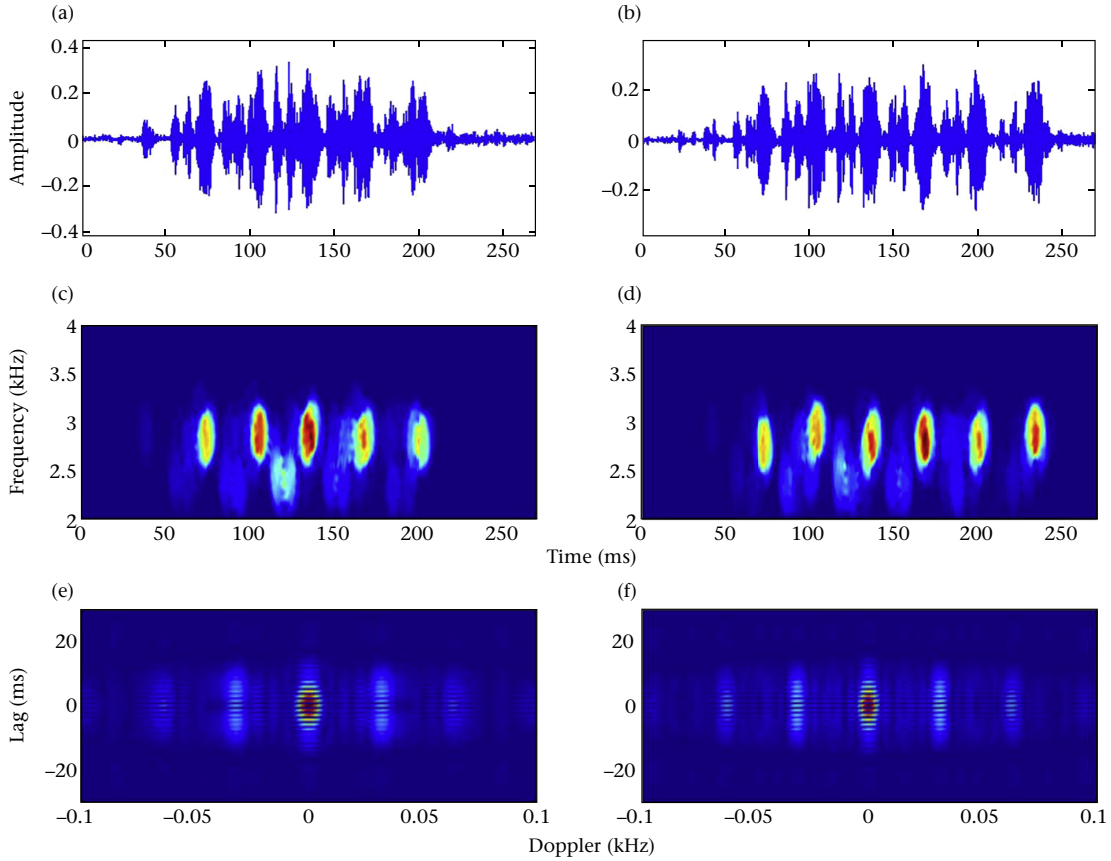


Figure 3. (a, b) Another example of two similar syllables. (a) Syllable C; (b) syllable D. (c, d) The corresponding multitaper spectrograms (linear scale) with time and frequency width 13.4 ms and 883 Hz and $M = 8$ windows for (c) syllable C and (d) syllable D. (e, f) The corresponding filtered ambiguity spectra of (e) syllable C and (f) syllable D.

specified beforehand (see Hastie, Tibshirani, & Friedman, 2009; Manning, Raghavan, & Schütze, 2008; Micheli-Tzanakou, 2000). Syllable-based birdsong analysis is a typical case of a clustering problem as the syllable groups are generally not known in advance, and nor are the number of groups into which the syllables can be partitioned. Among numerous available clustering algorithms, the hierarchical clustering approach is one that does not require the knowledge of the number of clusters and is therefore chosen for our application.

Hierarchical clustering

Hierarchical methods return a ranked structure, which is generally more informative than a fixed output of clusters. The partitions obtained by a hierarchical clustering method can be visualized by a dendrogram, a tree structure plot, which shows how the groups are nested at different levels (see e.g. Fig. 11a in example 3 of the section on evaluation and results below). Using different cutoff thresholds for pruning the cluster tree, the user can refine or coarsen the clusters and therefore achieve different levels of granularities. This is especially useful in application to birdsong analysis as researchers may choose the level of resolution depending on whether a fine-grained clustering is desirable or a grouping with less attention to details is sufficient. The interested reader may learn about the advantages and disadvantages of different hierarchical clustering algorithms in Manning et al. (2008). In our study we apply the agglomerative approach (i.e. the algorithm starts with defining each data point as its own cluster and successively merges clusters until only a single cluster, containing all data, remains) and similarity between clusters is calculated as the average of all pairwise similarities between members of the two clusters (the so-called average link method).

Optimal thresholding

To obtain a final partition of syllables into clusters, a cutoff threshold ρ for the syllable dendrogram has to be selected. The choice depends on the required resolution and deciding on an appropriate value of ρ (or, equivalently, for the number K of clusters) is generally not obvious. The literature proposes a variety of approaches for 'optimal' threshold selection, most of which are based on measuring the quality of a clustering for different values of K by means of internal or external quality criteria and the most common criteria for cluster evaluation are described in Maimon and Rokach (2005) and Zaki and Meira (2014).

External quality criteria use additional information that cannot be derived from the data, for instance external expert knowledge. This, however, is not always available. Internal quality criteria, in contrast, are exclusively based on information inherent in the data. They assess the quality of a clustering by measuring intracluster (which should be small) and intercluster variability (which is supposed to be large compared to intercluster spread). A neat and informative overview of most internal quality indices, their definition, references and implementations in the software R (R Core Team, 2014) can be found in Charrad, Ghazzali, Boiteau, and Niknafs (2014). In several comparative studies (see e.g. Arbelaitz,

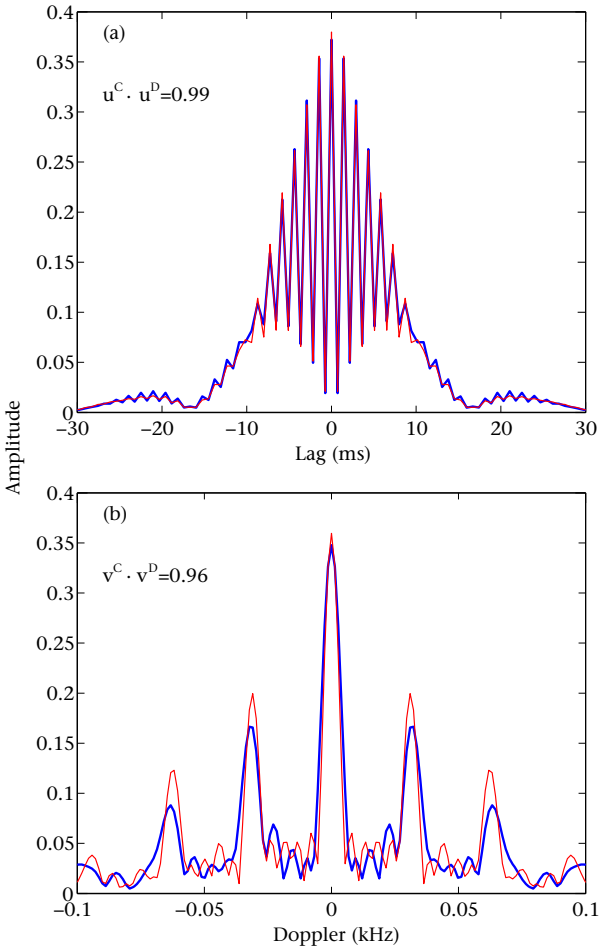


Figure 4. The corresponding (a) left and (b) right singular vectors of the filtered ambiguity spectra of syllable C (blue) and syllable D (red) shown in Fig. 3.

Gurrutxaga, Muguerza, Pérez, & Perona, 2013; Guerra, Robles, Bielza, & Larranga, 2012) the Silhouette quality measure (Rousseeuw, 1987; Zaki & Meira, 2014) was found to perform well in our context of hierarchical clustering with average link and is the criterion of choice in the following. The Silhouette index takes values between −1 and 1 and the higher the Silhouette value, the better is the clustering. In particular, this criterion does not require any subjective decisions by the user such as finding a 'knee' in a plot or a tuning of parameters. It therefore facilitates our goal to provide a fully objective approach for birdsong analysis.

EVALUATION AND RESULTS

In this section, we assess the performance of our proposed filtered ambiguity MT spectrum algorithm when applied to real data and compare it to both an MFCC-based method and to a method relying on SPCC. Thus, this section presents three examples that serve as a demonstration of the method's applicability.

Example 1: Separating two Syllable Classes

In this first example the methods are applied to classify a set of 39 syllables into two classes, which can easily be distinguished by eye and ear (Fig. 5). The within-class variability of the syllables is, however, very high (suggesting several subclasses, see example 2), which provides an extra challenge for the algorithms. Class 1 contains 17 well-aligned similar syllables, although of somewhat different amplitudes. The two syllables marked A and B are those exemplified in Fig. 2. Class 2 consists of 22 syllables, which are similar in structure, but not necessarily in time-frequency patterns. The two syllables marked C and D are exemplified in Fig. 3. The syllables of each class are time aligned using pairwise time correlation.

To put the performance of our technique into perspective, we additionally compare the results of our approach to those one would obtain if a single Hanning window spectrogram were used for calculation of the filtered ambiguity spectrum. All methods are evaluated in terms of a receiver operating characteristics (ROC) curve. The resolutions of the spectrogram-based methods are chosen optimally and are the same as those in Figs 2 and 3, i.e. 13.4 ms and 883 Hz for the MT spectrogram using $M = 8$ windows and 4.17 ms and 474 Hz for the single Hanning window spectrogram. For the MFCC method, the often used implementation by Slaney (1998) is chosen with eight cepstral coefficients, a 25 ms Hamming window and 90% overlap between frames. For the SPCC method the single window Hanning spectrogram with time and frequency resolutions of 4.17 ms and 474 Hz as defined above is employed. For the method-specific similarity measures, different thresholds are applied to find the ROC curves for all methods. The results are shown in Fig. 6 with the true positive rate (correctly classified as similar) on the *y*-axis and the false positive rate (erroneously classified as similar) on the *x*-axis.

The filtered ambiguity MT spectrum (blue line) clearly outperforms the other three methods with 90% correct classifications, accepting 5% false positives (Fig. 6). In contrast, the filtered ambiguity spectrum based on the noise-sensitive single Hanning window spectrogram (cyan line) achieves only 80% correct classifications (accepting 5% false positives). The well-known MFCC method (red line) and the SPCC approach (green line) fail markedly with only about 60% and 55% correct classifications, respectively.

Example 2: Clustering of the Syllable Set in Example 1

Typically, the researcher is faced with a large number of syllables and the question of how many different syllable types they represent. In this view, the setting of the first example with an a priori known number of classes is too simplistic for many applications. We now examine how a hierarchical clustering algorithm groups the syllable set from example 1, in which the number of clusters is determined via the Silhouette statistic.

The blue line in Fig. 7 shows the Silhouette values evaluated for a range of possible threshold values (between 0.033 and 0.2). The black numbers on top of the line denote the number of clusters corresponding to each threshold. The Silhouette statistic is in fact maximized for all threshold values within the interval [0.0620, 0.1130], all of them yielding the same clustering result. The dendrogram for this data set is depicted in Fig. 8. Applying (one of) the optimal threshold(s), the resulting first (purple) cluster contains syllables 1 up to 17 and therefore corresponds exactly to Class 1 in the manual classification. However, optimal thresholding splits Class 2 into four distinct groups, as illustrated in Fig. 9 (and denoted by Cluster 2 to Cluster 5). Examples of syllables from each of these four groups are shown in Appendix Fig. A1, using more common grey-scale spectrograms. For evaluation of this result a human expert (D.H.) conducted clustering of this syllable set. The expert-based result not only yielded the same number of clusters, the separation of the 39 syllables into the four groups was also identical to the algorithm-based classification.

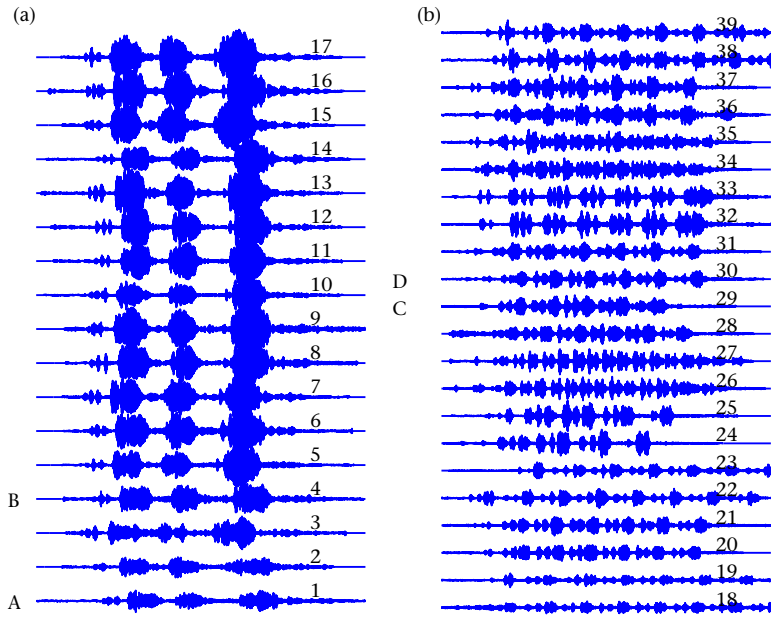


Figure 5. Two classes of syllables chosen from one song. (a) Class 1, 17 syllables; (b) Class 2, 22 syllables. Syllables A, B, C and D from Figs 2 and 3 are indicated.

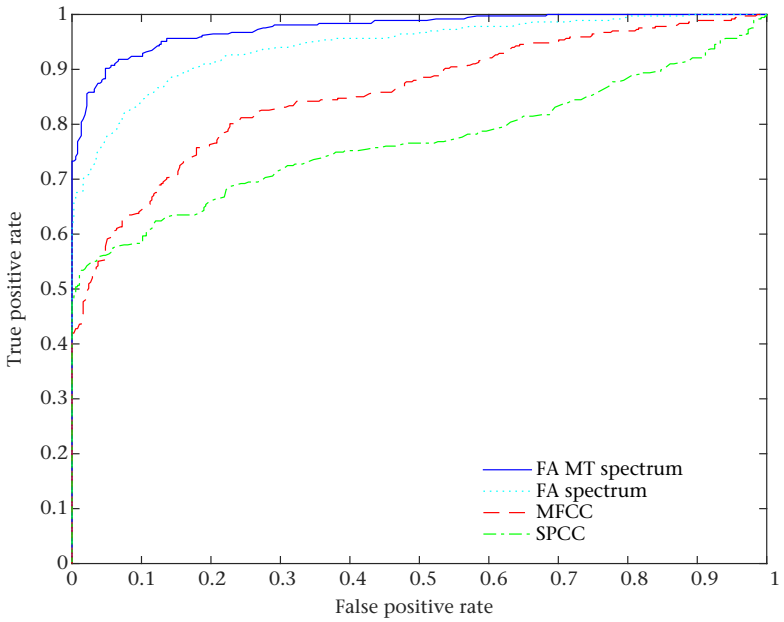


Figure 6. Receiver operating characteristics curves for the filtered ambiguity (FA) MT spectrum, the filtered ambiguity spectrum, the MFCC and the SPCC method.

Example 3: Clustering of a Whole Song

As a third application, we extend the previous clustering problem to a more complex setting of clustering the syllables of a whole song, which illustrates an interesting, nontrivial application of our method. The data are from a nearly 4 min field recording of a song of a GRW, comprising 433 syllables. As previously, we evaluate the Silhouette criterion for a range of threshold values, determine the value that maximizes the Silhouette statistic and use this threshold as the optimal cutoff for the dendrogram.

The Silhouette statistic, calculated for threshold values within the interval [0.033, 0.1], is displayed in Fig. 10a with a black circle marking the optimal threshold $\rho_{opt} = 0.0450$. The dendrogram from the hierarchical clustering algorithm is shown in Fig. 11a. Different colours illustrate the resulting syllable clusters and the black dashed horizontal line marks the optimal threshold. If more

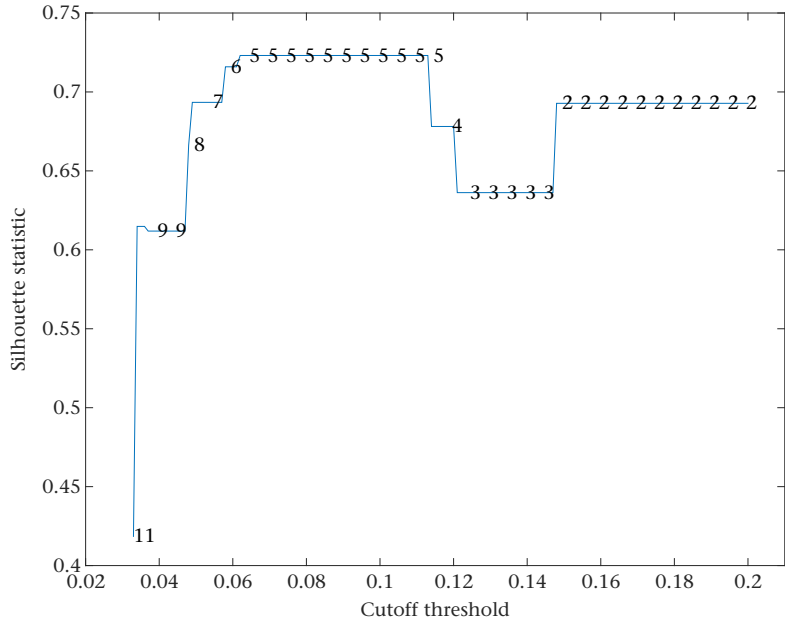


Figure 7. Investigation of the Silhouette statistic (the quality index) as a function of the cutoff threshold. The numbers denote the number of resulting clusters for the different thresholds.

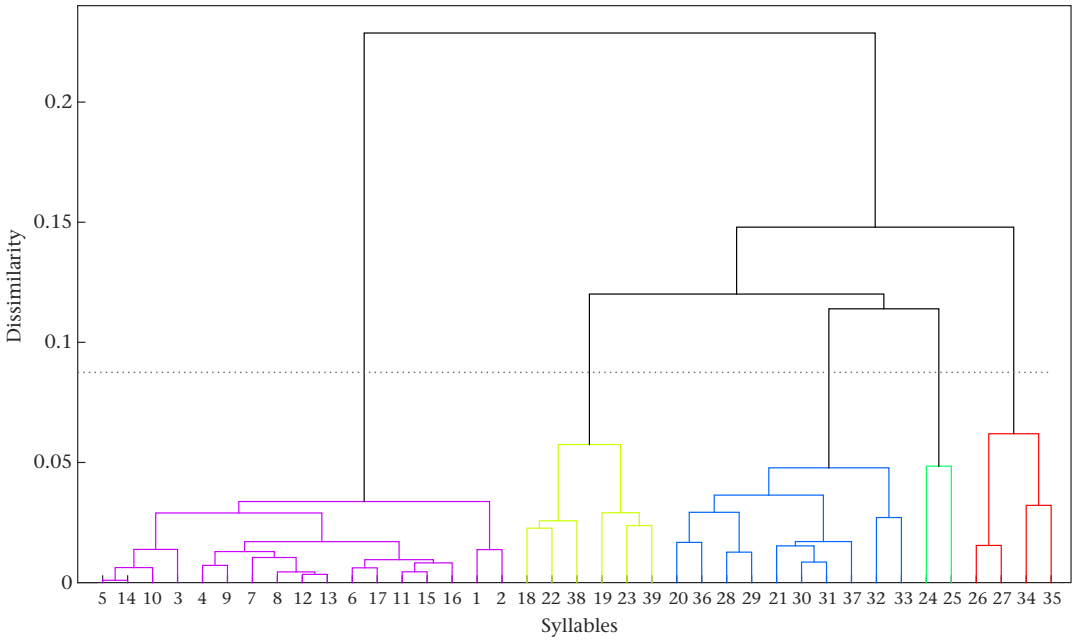


Figure 8. Dendrogram for the syllable set considered in the first two examples. The numbers on the *x*-axis correspond to the number of the syllable (as it occurs in the song) and the *y*-axis quantifies the dissimilarity between clusters (as calculated by the average link method, as described in the Methods section.). The horizontal dotted line represents the mean of the optimal thresholds.

resolution is desired, the user may want to lower the threshold and thus obtain more clusters. The way the number of clusters is related to the different thresholds can be seen in Fig. 10b. The chosen cutoff corresponds to grouping the syllables into 57 distinct clusters. For this clustering choice, the median number of syllables in each cluster is 7 and the largest cluster contains 35 syllables. It should, however, be noted that the Silhouette statistic can only be a guidance for the selection of a threshold. As can be seen in Fig. 10, the Silhouette optimum is not very pronounced, implying that clustering results for other threshold choices are nearly as good (when

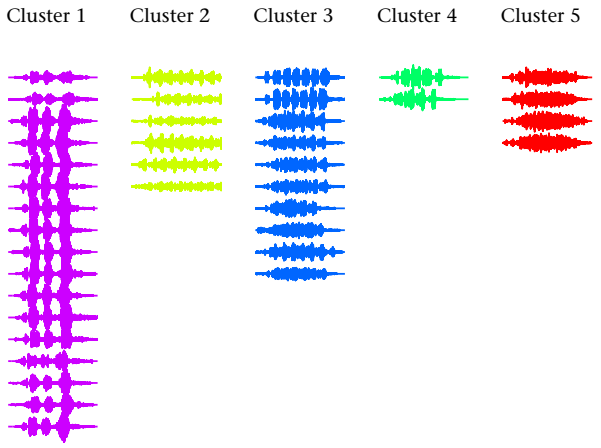


Figure 9. Syllable clusters. Cluster 1 coincides with Class 1 from example 1. The remaining syllable columns constitute Class 2 but are here (using the optimal threshold) grouped into four distinct classes.

quality is measured in terms of the Silhouette statistic). It is always recommendable to conduct a manual postanalysis and compare clusterings for different thresholds. This comparison is facilitated by the dendrogram, which easily visualizes how clusterings change depending on the chosen thresholds. Figure 11b shows a zoom-in of the dendrogram in Fig. 11a and the syllables occurring in the three different clusters highlighted there are displayed in Fig. 11c. Syllables within the same cluster are in fact (visually) very similar, while being rather distinct between different clusters. To further investigate the clustering obtained by the Silhouette-optimal threshold choice, the 433 syllables have also been classified by a human expert, based on visual inspection, which resulted in 37 distinct clusters. The analysis in example 2 was conducted on a smaller part of the data considered here. It is therefore interesting how syllables are clustered when the optimal threshold is estimated based on only a partial data set. Applying the threshold $\rho = 0.0620$, which was an optimal choice in Example 2 and, moreover, resulted in a clustering that exactly coincided with the result obtained by a human expert, gives a separation of the 433 syllables into a somewhat coarser clustering with 45 distinct clusters.

DISCUSSION

There is a need for a noise-insensitive automated analysis tool that allows detailed investigations of more complex birdsongs recorded under natural outdoor conditions. This is because standard previous approaches are highly time consuming and subjective (manual visual/audial inspection) or very sensitive to sound noise and time/frequency shifts (spectrogram-based, single-domain cross-correlations). The unavailability of reliable and fast birdsong analysis tools has hampered research on birdsong, in particular in wild species with moderate to high song complexity. Hence, much of what we know today about birdsong is based on information from species with low variation in their song. This can bias our understanding of the meaning and functions of birdsong, as well as the evolution of song repertoires and song complexity in general (Catchpole, 1989; Kroodsma, 1989).

The method we propose here serves as a new tool for the objective analysis of complex birdsongs. It is worth mentioning, though, that the performance of an algorithm for birdsong analysis depends heavily on the way syllables are determined (e.g. the tricky

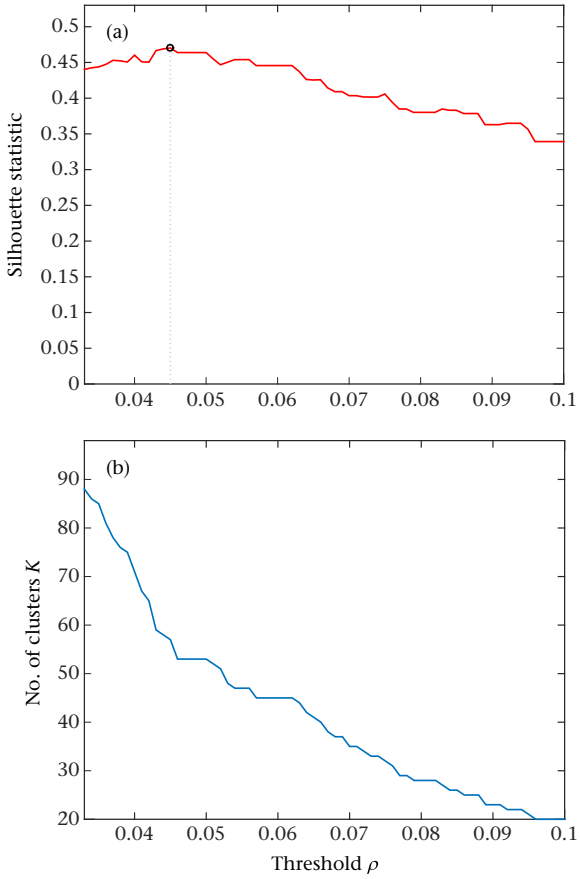


Figure 10. (a) Evaluation of the Silhouette statistic (the quality index) as a function of the threshold ρ. The optimal threshold ρ_{opt} and the corresponding Silhouette value are marked as a black circle. (b) Number K of clusters as a function of the threshold value ρ.

issue of whether closely spaced syllables are considered to be two single syllables or one double syllable), on the selection of representing features and on the similarity measure. Therefore, results always have to be viewed in the light of the (subjectively) chosen syllable representation and classification method. However, a fully automated analysis procedure does not add further subjectivity to the song investigation. Therefore, after a researcher has decided on an automated analysis procedure, the results obtained are not influenced by subjectivity, ensuring reproducibility and comparability.

To provide an objective method of threshold determination, we evaluate the quality of a clustering in terms of the Silhouette index. This criterion favours clusterings with a Silhouette-optimal trade-off between high intracluster similarity and high intercluster differences, offering reasonable guidance for the threshold choice. One should, however, always keep in mind that a Silhouette-optimal clustering result can never be more than an initial guide. A Silhouette-optimal threshold might not coincide with the desired coarseness of the clustering and therefore other, nearly optimal, thresholds might fit the specific research question better. Investigations that focus on song-based species recognition, for instance, may require a lower level of resolution than within-individual song analyses. Moreover, one should bear in mind that the clustering algorithm itself will always be prone to misclassifications, independent of the final choice of the cutoff level.

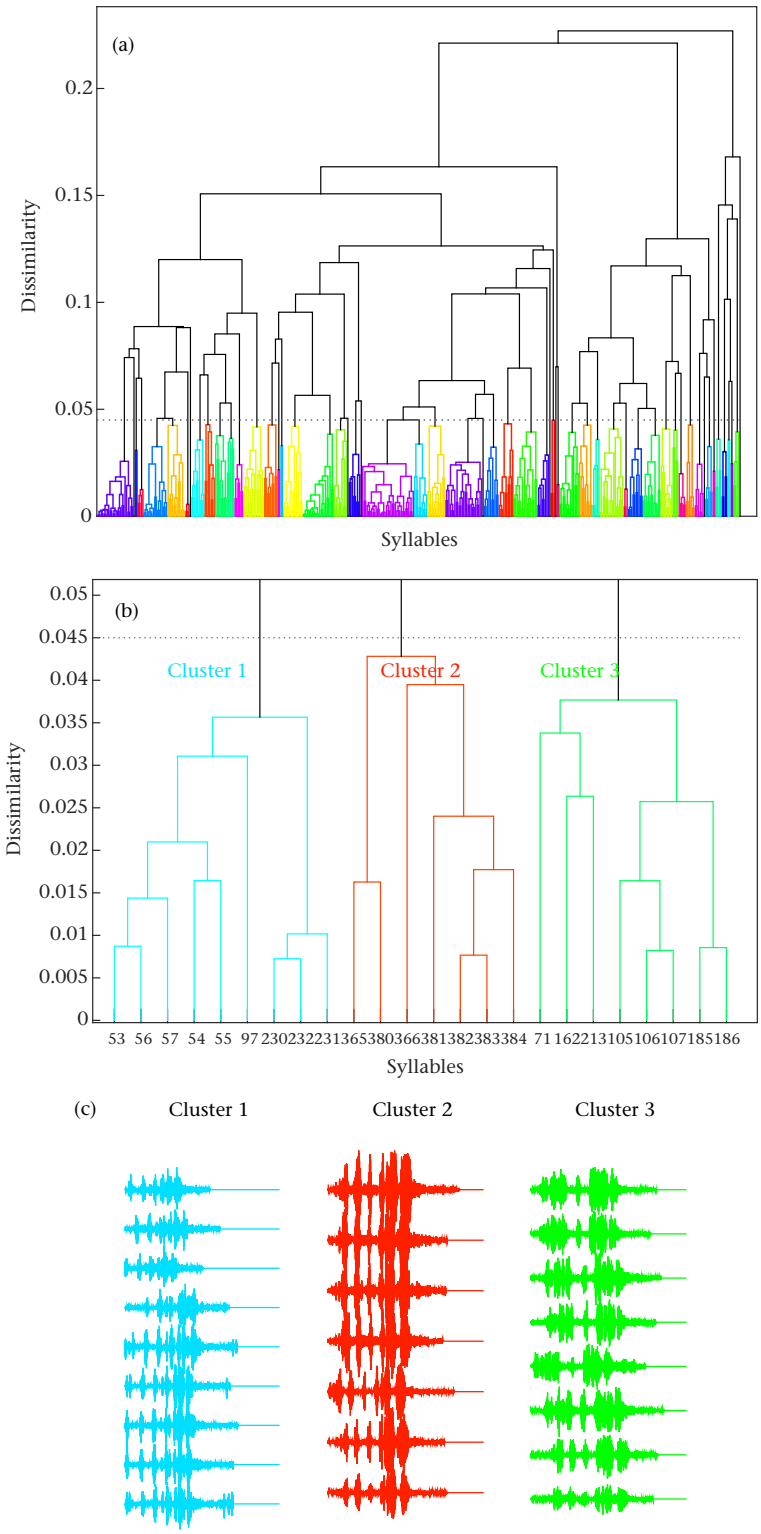


Figure 11. Clustering result for example 3. (a) Illustration of the hierarchical clustering result as a dendrogram. The horizontal dotted line marks the cutoff at the optimal threshold $\rho_{opt} = 0.0450$. (b) Zoom-in of the dendrogram in (a). (c) The syllables in the three clusters from (b).

The researcher is therefore always advised to conduct a manual postanalysis by inspecting the clusters in more detail (e.g. by going back to the time and/or the time-frequency representation of the syllables) and to thereby correct possible misclassifications of the algorithm. However, despite the need for manual inspection/adjustment, an automated preclustering will save researchers a considerable amount of time and will serve as a useful guide in the analysis of birdsong. For situations in which a researcher is not interested in a final clustering or dendrogram, but solely in the pairwise similarity scores, the algorithm can be adjusted to return only the matrix of similarity scores. The novelty in our approach is the combination of multitapering and the ambiguity spectrum, including SVD-based feature extraction. The chosen syllable representation, along with the specific similarity measure, captures key properties of birdsong syllables and thereby leads to a powerful algorithm. The method, applied to real data, is evaluated in terms of comparison to human expert evaluation and gives very promising results in the context of clustering of GRW syllables. It, moreover, clearly outperforms existing approaches based on MFCC or SPCC. Additional improvements could be achieved by (1) expanding the chosen feature set by some additional features and (2) weighting features according to their importance. In this way a syllable s would be represented by a feature matrix $F = [u, v, b]$, where u, v are the first left and right singular vectors obtained from the SVD of the estimated filtered ambiguity matrix and b contains additional features, such as the time length of a unit, its power or pitch frequency. If $\alpha(\cdot, \cdot)$ is another similarity measure suitable for assessing the similarity of two vectors $b^{(i)}$, $b^{(j)}$ (for instance in terms of the Euclidean distance), one could investigate alikeness of two syllables $s^{(i)}$, $s^{(j)}$ by means of the (weighted) similarity measure $\gamma(s^{(i)}, s^{(j)}) = \lambda\beta(s^{(i)}, s^{(j)}) + (1 - \lambda)\alpha(s^{(i)}, s^{(j)})$ for some λ between 0 and 1.

Acknowledgments

This work was supported by the Swedish strategic research programme eSSENCE (grants to M.S.), the Swedish Research Council (grants to D.H., B.H.), Research Council of Norway through its Centres of Excellence funding scheme (project number 223257, M.T.), Lunds Djurskyddsfond (to M.T., B.H., D.H.), CAnMove (a Linnaeus research excellence environment financed by Swedish Research Council and Lund University) and Kvismare Bird Observatory (177). Moreover, we thank Jessica Caissy-Martineau for evaluation of earlier versions of the song analysis program.

References

Adret, P., Meliza, C. D., & Margoliash, D. (2012). Song tutoring in presinging zebra finch juveniles biases a small population of higher-order song-selective neurons toward the tutor song. *Journal of Neurophysiology, 108*(7), 1977–1987.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition, 46*(1), 243–256.

Bayram, M., & Baraniuk, R. G. (1996). Multiple window time-frequency analysis. *Proceedings of the International Symposium of Time-Frequency and Time-Scale Analysis*, 511–514.

Boashash, B. (2003). Theory of quadratic TFDs. In B. Boashash (Ed.), *Time frequency signal analysis and processing: A comprehensive reference* (pp. 59–82). Oxford, U.K.: Elsevier.

Boersma, P., & Weenink, D. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*(9/10), 341–345.

Bronez, T. P. (1992). On the performance advantage of multitaper spectral analysis. *IEEE Transactions on Signal Processing, 40*(12), 2941–2946.

Catchpole, C. K. (1976). Temporal and sequential organisation of song in the sedge warbler (*Acrocephalus schoenobaenus*). *Behaviour, 59*(3), 226–245.

Catchpole, C. K. (1983). Variation in the song of the great reed warbler *Acrocephalus arundinaceus* in relation to mate attraction and territorial defence. *Animal Behaviour, 31*(4), 1217–1225.

Catchpole, C. K. (1989). Pseudoreplication and external validity: some thoughts on the suggested redesign of playback experiments in avian bioacoustics. *Trends in Ecology and Evolution, 4*, 286–287.

Catchpole, C. K., & Rowell, A. (1993). Song sharing and local dialects in a population of the European wren *Troglodytes troglodytes*. *Behaviour, 125*(1), 67–78.

Catchpole, C. K., & Slater, P. J. (1995). *Bird song: Biological themes and variations*. Cambridge, U.K.: Cambridge University Press.

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). Nbclust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software, 61*(6), 1–36.

Clark, C. W., Marler, P., & Beeman, K. (1987). Quantitative analysis of animal vocal phonology: an application to swamp sparrow song. *Ethology, 76*(2), 101–115.

Daubechies, I. (1988). Time-frequency localization operators: a geometric phase space approach. *IEEE Transactions on Information Theory, 34*(4), 605–612.

Eriksen, A., Slagsvold, T., & Lampe, H. M. (2011). Vocal plasticity – are pied flycatchers, *Ficedula hypoleuca*, open-ended learners? *Ethology, 117*(3), 188–198.

Espmark, Y. O., Lampe, H. M., & Bjerke, T. K. (1989). Song conformity and continuity in song dialects of redwings *Turdus iliacus* and some ecological correlates. *Ornis Scandinavica*, 1–12.

Fagerlund, S. (2007). Bird species recognition using support vector machines. *EURASIP Journal on Applied Signal Processing, 2007*(1), 64.

Falls, J. B. (1985). Song matching in western meadowlarks. *Canadian Journal of Zoology, 63*(11), 2520–2524.

Guerra, L., Robles, V., Bielza, C., & Larranga, P. (2012). A comparison of clustering quality indices using outliers and noise. *Intelligent Data Analysis, 16*(4), 703–715.

Hansson-Sandsten, M. (2011). Optimal multitaper wigner spectrum estimation of a class of locally stationary processes using hermite functions. *EURASIP Journal on Advances in Signal Processing, 2011*, 10.

Härmä, A. (2003, April). Automatic identification of bird species based on sinusoidal modeling of syllables. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03). 2003 IEEE International Conference on* (Vol. 5, pp. V–545). IEEE.

Hasselquist, D. (1998). Polygyny in great reed warblers: a long-term study of factors contributing to male fitness. *Ecology, 79*(7), 2376–2390.

Hasselquist, D., & Bensch, S. (1991). Trade-off between mate guarding and mate attraction in the polygynous great reed warbler. *Behavioral Ecology and Sociobiology, 28*(3), 187.

Hasselquist, D., Bensch, S., & von Schantz, T. (1996). Correlation between male song repertoire, extra-pair paternity and offspring survival in the great reed warbler. *Nature, 381*, 229–232.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (Vol. 2). Berlin, Germany: Springer.

Horn, A. G., & Falls, J. B. (1988). Repertoires and countersinging in western meadowlarks (*Sturnella neglecta*). *Ethology, 77*(4), 337–343.

Horn, A. G., & Falls, J. B. (1996). Categorization and the design of signals: the case of song repertoires. In D. E. Kroodsma, & E. H. Miller (Eds.), *Ecology and evolution of acoustic communication in birds* (pp. 121–135). Ithaca, NY: Cornell University Press.

Keen, S., Ross, J. C., Griffiths, E. T., Lanzone, M., & Farnsworth, A. (2014). A comparison of similarity-based approaches in the classification of flight calls of four species of North American wood-warblers (*Parulidae*). *Ecological Informatics, 21*, 25–33.

Kogan, J. A., & Margoliash, D. (1998). Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: a comparative study. *The Journal of the Acoustical Society of America, 103*(4), 2185–2196.

Kreutzer, M., & Güttinger, H. R. (1991). Konkurrenzbeziehungen und Verhaltensantworten gegenüber dem Gesang: Artnorm und individuelle Variabilität bei der Zaunammer (*Emberiza cirlus*). *Journal für Ornithologie, 132*(2), 165–177.

Kroodsma, D. E. (1989). Suggested experimental designs for song playbacks. *Animal Behaviour, 37*, 600–609.

Kroodsma, D. E., & Konishi, M. (1991). A suboscine bird (Eastern phoebe, *Sayornis phoebe*) develops normal song without auditory feedback. *Animal Behaviour, 42*(3), 477–487.

Lachlan, R. F. (2007). *Luscinia: A bioacoustics analysis computer program*. Retrieved from www.lusciniasound.org.

Lampe, H. M., & Espmark, Y. O. (2003). Mate choice in pied flycatchers *Ficedula hypoleuca*: can females use song to find high-quality males and territories? *Ibis, 145*(1), E24–E33.

Lehtonen, L. (1983). The changing song patterns of the great tit *Parus major*. *Ornis Fenn, 60*, 16–21.

Lemon, R. E., Cotter, R., MacNally, R. C., & Monette, S. (1985). Song repertoires and song sharing by American redstarts. *Condor*, 457–470.

Maimon, O., & Rokach, L. (Eds.). (2005). *Data mining and knowledge discovery handbook* (Vol. 2). New York, NY: Springer.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge, U.K.: Cambridge University Press.

Martens, J. (1996). Vocalizations and speciation of palearctic birds. In D. E. Kroodsma, & E. H. Miller (Eds.), *Ecology and evolution of acoustic communication in birds* (pp. 221–240). Ithaca, NY: Cornell University Press.

McGregor, P. K. (1980). Song dialects in the corn bunting (*Emberiza calandra*). *Zeitschrift für Tierpsychologie, 54*(3), 285–297.

Meliza, C. D., Keen, S. C., & Rubenstein, D. R. (2013). Pitch-and spectral-based dynamic time warping methods for comparing field recordings of harmonic avian vocalizations. *The Journal of the Acoustical Society of America, 134*(2), 1407–1415.

Micheli-Tzanakou, E. (Ed.). (2000). *Supervised and unsupervised pattern recognition: Feature extraction and computational intelligence*. Boca Raton, FL: CRC Press.
Miller, E. H. (1996). Acoustic differentiation and speciation in shorebirds. In D. E. Kroodsma, & E. H. Miller (Eds.), *Ecology and evolution of acoustic communication in birds* (pp. 241–257). Ithaca, NY: Cornell University Press.
Miller, E. H., & Kroodsma, D. E. (Eds.). (1996). *Ecology and evolution of acoustic communication in birds*. Ithaca, NY: Cornell University Press.
Mundinger, P. C. (1980). Animal cultures and a general theory of cultural evolution. *Ethology and Sociobiology, 1*(3), 183–223.
Nottebohm, F. (1991). Reassessing the mechanisms and origins of vocal learning in birds. *Trends in Neurosciences, 14*(5), 206–211.
R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.r-project.org.
Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65.
Sandsten, M., Tarka, M., Caissy-Martineau, J., Hansson, B., & Hasselquist, D. (2011). A SVD-based classification of bird singing in different time-frequency domains using multitapers. In *19th European SIgnal Processing Conference (EUSIPCO-2011)* (Vol. 2011, pp. 966–970). European Association for Signal Processing (EURASIP).
Searcy, W. A., & Yasukawa, K. (1996). Song and female choice. In D. E. Kroodsma, & E. H. Miller (Eds.), *Ecology and evolution of acoustic communication in birds* (pp. 454–473). Ithaca, NY: Cornell University Press.
Selin, A., Turunen, J., & Tanttu, J. T. (2007). Wavelets in recognition of bird sounds. *EURASIP Journal on Applied Signal Processing, 2007*(1), 141.
Slaney, M. (1998). *Auditory toolbox: Version 2* (Technical Report 1998-010). Interval Research Corporation. Retrieved from https://engineering.purdue.edu/~malcolm/interval/1998-010.
Slater, P. J. B. (1983). Sequences of song in chaffinches. *Animal Behaviour, 31*(1), 272–278.
Slater, P. J. B., Clements, F. A., & Goodfellow, D. J. (1984). Local and regional variations in chaffinch song and the question of dialects. *Behaviour, 88*(1), 76–97.
Slater, P. J. B., & Ince, S. A. (1982). Song development in chaffinches: what is learnt and when? *Ibis, 124*(1), 21–26.
Somervuo, P., & Härmä, A. (2004, May). Bird song recognition based on syllable pair histograms. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP'04). IEEE International Conference on* (Vol. 5)IEEE. pp. V–825.
Somervuo, P., Härmä, A., & Fagerlund, S. (2006). Parametric representations of bird sounds for automatic species recognition. *IEEE Transactions on Audio, Speech, and Language Processing, 14*(6), 2252–2263.
Specht, R. (2004). *Avisoft-SASLab Pro*. Berlin, Germany: Avisoft.
Tchernichovski, O., Lints, T., Mitra, P. P., & Nottebohm, F. (1999). Vocal imitation in zebra finches is inversely related to model abundance. *Proceedings of the National Academy of Sciences of the United States of America, 96*(22), 12901–12904.
Tchernichovski, O., & Mitra, P. P. (2004). *Sound analysis pro user manual*. Retrieved from http://soundanalysispro.com/manual-1.
Tchernichovski, O., Nottebohm, F., Ho, C. E., Pesaran, B., & Mitra, P. P. (2000). A procedure for an automated measurement of song similarity. *Animal Behaviour, 59*(6), 1167–1176.
Thomson, D. J. (1982). Spectrum estimation and harmonic analysis. *Proceedings of the IEEE, 70*(9), 1055–1096.
Trifa, V. M., Kirschel, A. N., Taylor, C. E., & Vallejo, E. E. (2008). Automated species recognition of antbirds in a Mexican rainforest using hidden Markov models. *The Journal of the Acoustical Society of America, 123*(4), 2424–2431.
Węgrzyn, E., & Leniowski, K. (2010). Syllable sharing and changes in syllable repertoire size and composition within and between years in the great reed warbler, *Acrocephalus arundinaceus*. *Journal of Ornithology, 151*(2), 255–267.
Węgrzyn, E., Leniowski, K., & Osiejuk, T. S. (2010). Whistle duration and consistency reflect philopatry and harem size in great reed warblers. *Animal Behaviour, 79*(6), 1363–1372.
Welch, P. D. (1967). The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics, 15*(2), 70–73.
Williams, J. M. (1993). Objective comparisons of song syllables: a dynamic programming approach. *Journal of Theoretical Biology, 161*(3), 317–328.
Williams, J. M., & Slater, P. J. B. (1991). Computer analysis of bird sounds: a guide to current methods. *Bioacoustics, 3*(2), 121–128.
Xu, Y., Haykin, S., & Racine, R. J. (1999). Multiple window time-frequency distribution and coherence of EEG using Slepian sequences and Hermite functions. *IEEE Transactions on Biomedical Engineering, 46*(7), 861–866.
Zaki, M. J., & Meira, W., Jr. (2014). *Data mining and analysis: Fundamental concepts and algorithms*. Cambridge, U.K.: Cambridge University Press.

Appendix

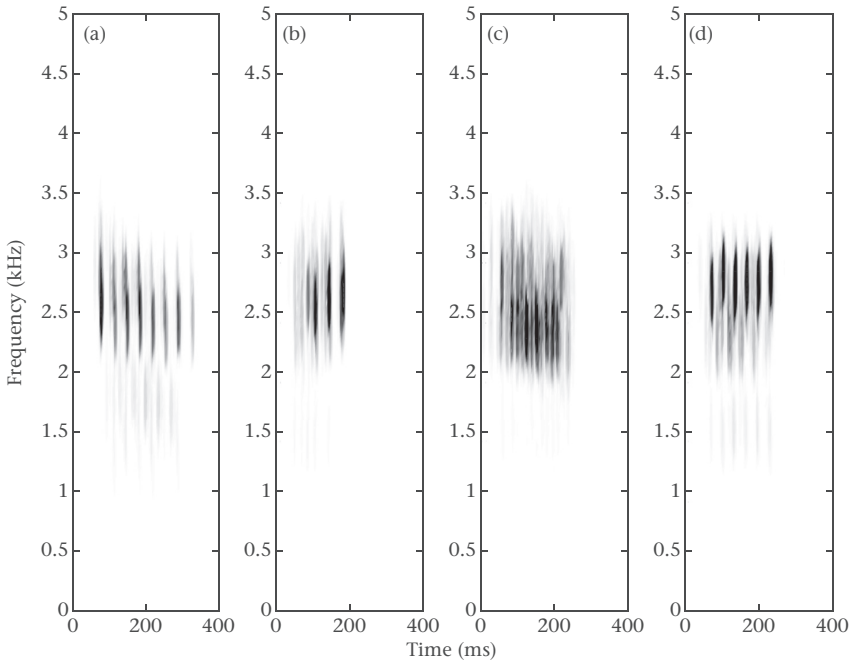


Figure A1. Spectrogram of four syllables, labels from Fig. 5, one from each of clusters 2–5 in Fig. 9. (a) Syllable 23; (b) syllable 24; (c) syllable 26; (d) syllable 30.

II - Absorption and initial metabolism of ^{75}Se-L-selenomethionine: a kinetic model based on dynamic scintigraphic data

Published in *British Journal of Nutrition* (2015), **114** (10), 1718-1723
DOI: 10.1017/S000711451500344X

Mareile Große Ruse
Department of Mathematical Sciences
University of Copenhagen, Denmark

L.R. Søndergaard
Department of Clinical Physiology and Nuclear Medicine,
Centre of Functional Imaging and Research
Hvidovre Hospital, Hvidovre, Denmark

Susanne Ditlevsen
Department of Mathematical Sciences
University of Copenhagen, Copenhagen, Denmark

Morten Damgaard
Department of Clinical Physiology and Nuclear Medicine
Centre of Functional Imaging and Research
Hvidovre Hospital, Hvidovre, Denmark

Stefan Fuglsang
Department of Clinical Physiology and Nuclear Medicine
Centre of Functional Imaging and Research
Hvidovre Hospital, Hvidovre, Denmark

Jimmy Ottesen
Department of Science, Systems and Models
Roskilde University, Roskilde, Denmark

Jan Lysgård Madsen
Department of Clinical Physiology and Nuclear Medicine
Centre of Functional Imaging and Research
Hvidovre Hospital, Hvidovre, Denmark

Keywords: Selenomethionine, ^{75}Se-L-selenomethionine, Absorption capacity, Metabolism, Gamma camera imaging, Compartmental modelling

British Journal of Nutrition (2015), **114**, 1718–1723
doi:10.1017/S000711451500344X

Absorption and initial metabolism of ^{75}Se-L-selenomethionine: a kinetic model based on dynamic scintigraphic data

Mareile Große Ruse[1], Lasse R. Søndergaard[2], Susanne Ditlevsen[1], Morten Damgaard[2], Stefan Fuglsang[2], Johnny T. Ottesen[3] and Jan L. Madsen[2]*

[1]*Department of Mathematical Sciences, Laboratory of Applied Statistics, University of Copenhagen, DK-2100 Copenhagen Ø, Denmark*
[2]*Department of Clinical Physiology and Nuclear Medicine, Centre of Functional Imaging and Research, Hvidovre Hospital, DK-2650 Hvidovre, Denmark*
[3]*Department of Science, Systems and Models, Roskilde University, DK-4000 Roskilde, Denmark*

(Submitted 10 March 2015 – Final revision received 30 July 2015 – Accepted 10 August 2015 – First published online 28 September 2015)

Abstract

Selenomethionine (SeMet) is an important organic nutritional source of Se, but the uptake and metabolism of SeMet are poorly characterised in humans. Dynamic gamma camera images of the abdominal region were acquired from eight healthy young men after the ingestion of radioactive ^{75}Se-L-SeMet (^{75}Se-SeMet). Scanning started simultaneously to the ingestion of ^{75}Se-SeMet and lasted 120 min. We generated time-activity curves from two-dimensional regions of interest in the stomach, small intestine and liver. During scanning, blood samples were collected at 10-min intervals to generate plasma time-activity curves. A four-compartment model, augmented with a delay between the liver and plasma, was fitted to individual participants' data. The mean rate constant for ^{75}Se-SeMet transport was 2·63 h^{-1} from the stomach to the small intestine, 13·2 h^{-1} from the small intestine to the liver, 0·261 h^{-1} from the liver to the plasma and 0·267 h^{-1} from the stomach to the plasma. The delay in the liver was 0·714 h. Gamma camera imaging provides data for use in compartmental modelling of ^{75}Se-SeMet absorption and metabolism in humans. In clinical settings, the obtained rate constants and the delay in the liver may be useful variables for quantifying reduced intestinal absorption capacity or liver function.

Key words: Selenomethionine: ^{75}Se-L-selenomethionine: Absorption capacity: Metabolism: Gamma camera imaging: Compartmental modelling

Selenomethionine (SeMet) is an important organic nutritional source of Se[1,2]. Absorption of various Se compounds occurs via different routes and mechanisms. Membrane transport of selenoamino acids, including SeMet, involves a specific suite of amino acid transporters[3]. The subsequent incorporation of dietary Se into selenoproteins occurs through a series of interconversions, of which many details remain unknown. Se metabolites are excreted in the urine and faeces and in exhaled air, mainly as selenosugars and methylated compounds[4].

The initial metabolism of Se in humans is poorly characterised. Estimates of Se absorption, whole-body retention and excretion have been made predominantly on whole-body counting[5] or the recovery of ingested tracers in the blood, urine and faeces[6]. Compartmental analyses of kinetic data from tracer studies have also been used to create a more integrated picture of whole-body Se utilisation in humans[7,8]. These studies characterised the long-term kinetics by the investigation of urine and faecal data collected over 12 d and blood samples drawn over 4 months. Through detailed mathematical modelling including several plasma pools, they were able to provide new insights into the long-run Se metabolism. However, because the study data only comprised hourly observations after dose administration, the initial Se kinetics could not be investigated and therefore still remained unclear. Our study tries to fill this gap and to provide deeper insight into the initial Se kinetics by focusing on frequent data collection within the first 2 h after administration. However, it should be noticed that the doses used in the previously mentioned studies[7,8] were considerably larger (150–200 μg) than those administered in the present study (29 μg), which might affect the kinetics and thus hamper the comparability of our study to the previous studies. In an earlier study[9], we had employed gamma camera imaging after oral intake of radio-labelled SeMet to quantify the gastrointestinal absorption capacity for SeMet and followed its postprandial distribution within the body. In the present study, we focused on dynamic gamma camera imaging with high temporal resolution to obtain data on both the intestinal absorption and the initial distribution

Abbreviations: ^{75}Se-L-SeMet, ^{75}Se-SeMet; ROI, region of interest; SeMet, selenomethionine.

* **Corresponding author:** J. L. Madsen, fax +45 3862 3750, email jan.lysgaard.madsen@regionh.dk

British Journal of Nutrition

British Journal of Nutrition

of SeMet in humans. We developed a compartmental model that was able to capture the behaviour of the high-resolution data and thus shed more light on the initial SeMet kinetics in humans. For the development of a suitable mathematical model, we followed two approaches. The first one used the simplest model, – the model with the fewest compartments and parameters – to explain the observed data by adding components to the model until an acceptable fit was achieved. In a second approach, we investigated the previously reported models[7,8] focusing solely on those model parts that corresponded to kinetics during the first 2 h after SeMet administration. Here, we subsequently eliminated terms until the parameters could be identified and an acceptable fit was achieved. Both approaches resulted in the same model.

Methods

Eight healthy men (age 24 (SD 3) years, weight 80·2 (SD 9·4) kg, height 1·81 (SD 0·05) m, BMI 24·6 (SD 3·0) kg/m^2 and plasma volume 3·37 (SD 0·23) litres) participated in the study. All participants exhibited normal plasma Se levels before commencement of the study (1·00 (SD 0·10) μmol/l). None of the participants had undergone previous abdominal surgery (other than appendectomy) or was receiving any medication. This study was conducted according to the guidelines laid down by the Declaration of Helsinki, and all procedures involving participants were approved by the scientific ethics committees of the Capital Region of Denmark (Protocol No. H-3-2009-092) and Danish Data Protection Agency (Journal No. 2009-41-3751). Written informed consent was obtained from all participants.

^{75}Se-L-selenomethionine

^{75}Se-L-SeMet (^{75}Se-SeMet) was produced and delivered by Hevesy Laboratory, DTU Nutec, Technical University of Denmark, Roskilde, Denmark, as described previously[9].

Procedure

Each participant arrived at the laboratory after having fasted for at least 10 h. A cannula was inserted into the cubital vein for blood sampling. Lying supine on the gamma camera couch, the participants then ingested 3·6 (SD 0·3) MBq of ^{75}Se-SeMet, comprising 29 μg Se dissolved in 350 ml of water. The solution was ingested in <15 s. The distribution of ^{75}Se-SeMet was investigated for the following 2 h using dynamic gamma camera imaging. Thus, 120 1-min images of the abdominal region were acquired in both anterior and posterior projections. Imaging was performed with a dual-head gamma camera equipped with medium-energy, all-purpose collimators (Infinia VC HawkEye; GE Medical Systems Inc.) and connected to a dedicated image processing system (Xeleris; GE Medical Systems Inc.). The images were acquired in a 128 × 128 matrix, with each pixel measuring 4·4 × 4·4 mm and using 136 keV (± 10 %) and 272 keV (± 12·5 %) energy windows.

During gamma camera imaging, 10-ml blood samples were collected at 10-min intervals to monitor the plasma concentration of ^{75}Se.

Processing of gamma camera data

To correct for ^{75}Se gamma ray attenuation caused by the gamma camera couch, a couch transmission factor was determined from an *in vitro* study with an approximated point source of 0·4 MBq of ^{75}Se placed in the centre of the gamma camera detection field, with one detector above (anterior) and one detector below (posterior) the couch. For both ^{75}Se energy windows, we found that the counts in a small region of interest (ROI) in the posterior-view image were about 90 % of the counts in the anterior-view image. In human studies, therefore, posterior-view counts were scaled up by a factor of 10/9 = 1·11. To compensate for gamma ray attenuation within participants' bodies, pixel-by-pixel geometric mean images were generated from conjugate anterior and adjusted posterior images. Finally, the geometric mean images were analysed for activity in the stomach, small intestine and liver using ROI delineated manually by the same observer. Because of the small number of counts in each 1 min image, it was necessary to summarise the images to obtain a resolution that permitted reliable delineation of the ROI. Hence, the images were summarised in periods over 0–30 min for drawing the stomach ROI and over 30–120 min for drawing the small intestine and liver ROI (Fig. 1).

Plasma analysis

Blood samples were centrifuged immediately for 10 min at 1000 ***g***, and the plasma was stored at −20°C until further analysis. To measure ^{75}Se activity, 3-ml aliquots of plasma and an appropriate dilution of stock solution of ^{75}Se-SeMet were counted for 30 min in a gamma well counter (Wizard 1480; Wallac Oy). For conversion into counts for total plasma volume, the total plasma volume of each participant was estimated from their height and weight data, according to tabulated references[10]. All counts were corrected for physical decay and expressed as a percentage of the administered activity.

Kinetic modelling

The kinetic model was developed to simultaneously capture the dynamics of data for stomach, liver, small intestine and plasma[11]. Parameter estimation and simulation of the compartmental models were carried out using Monolix[12]. The way in which data were collected gives rise to specific difficulties such as overlapping tissues. In Fig. 1, for instance, it is clearly

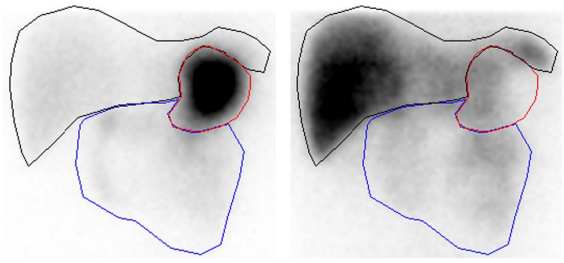


Fig. 1. Representative regions of interest for sampling of scintigraphic data (subject A): stomach (red), small intestine (blue) and liver (black). Left image summarised over 0–30 min. Right image summarised over 30–120 min.

British Journal of Nutrition

visible that the stomach ROI covered parts of both liver ROI and small intestine ROI. Moreover, all three organs had underlying blood flow that contributed to the counts. Both phenomena had to be accounted for by the model. Additionally, we incorporated uncertainties in the data collection procedure itself (observation error) into our model. Hence, the final model consists of three parts. The first stage describes the hidden states, assumed to be the actual tissues under study, but not directly observable due to overlapping tissues and measurement uncertainties. The second part consists of the observational states. Here the stomach counts contain additional contributions from the liver, small intestine and plasma, and the liver and small intestine contain – apart from the tissues themselves (except those fractions that were erroneously interpreted as stomach counts) – additional contributions from plasma. Plasma is assumed to be directly observed, as these data were obtained from blood samples. The third stage of the model is given by the observation equations, which model measurement uncertainties. In this last part, both additive and multiplicative errors as well as combinations thereof were considered. Moreover, a random subject-specific component was introduced that allowed model parameters to vary across subjects and, therefore, to account for inter-individual variations in the model. This was achieved by fitting the data from all subjects to one overall model and at the same time assuming that the model parameters were drawn from a population (the population of subjects). That is, across different subjects, the parameters were assumed to vary randomly around their respective median value (the population estimate), and the extent of variation (i.e. their variances) quantifed the variability among subjects. This modelling approach improves population estimates compared with averages of individual estimates. The specific distributions used for the random effects were also part of the model building. Several models were tried to describe the data until a suitable model was found that provided an adequate fit with no systematic deviations. The Akaike information criterion and Bayesian information criterion were used to choose among the models. These are measures of the relative quality of the considered statistical models for a given set of data and can be viewed as measures that combine the goodness of fit and the complexity of a model. Finally, models developed in previous studies[7,8] also were tried, even though they were developed for a different timescale and for other SeMet doses. However, neither these models nor similar versions thereof were able to describe the data. This might be attributed to the fact that those models were, on the one hand, initially developed for longer time scales and, on the other hand, were originally fitted to rather different types of data (plasma, urine and faeces) and were without data from the stomach, small intestine and liver. As these previous models seem to be inadequate to account for the present type of data (frequent recordings of initial SeMet distribution in the stomach, small intestine, liver and plasma), we opted for the most parsimonious model that was able to explain the data. The final model consists of four compartments, one for each of the observed tissues, including a delay between the liver and plasma. The dose arrived to the stomach with a short distributed delay, and all flows between the compartments were best modelled with first order kinetics. All random effects were best modelled with log-normal distributions, except for two parameters (a_L and k_e), which were better fitted with their square root being normally distributed. Finally, the observation error for all compartments had a multiplicative component, with an additional additive component for liver and stomach data. The model is illustrated in Fig. 2, and model equations can be found in Table 1.

For each participant, the activity in each compartment was normalised by the maximum value over time of the sum of all four compartments. This normalising value corresponds to the initial dose in counts, and numbers can be interpreted as percentage of initial dose.

The disappearance half-life of ^{75}Se-SeMet in the liver ($t_{1/2}$) was calculated using $t_{1/2} = \ln(2)/k_4$ where k_4 is the outflow of ^{75}Se-SeMet from the liver into the plasma.

Results

The measured time-activity curves for all participants are shown in Fig. 3. The black thick curve is the population fit of the model to data. An example of measured and fitted data from one individual is shown in Fig. 4.

All estimated values are given in Table 2. Thus, Table 2 shows the estimated rate parameters for the transport of ^{75}Se-SeMet between the compartments and further model

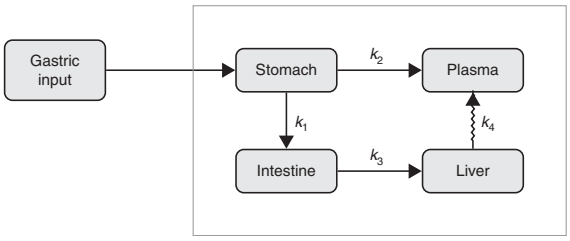


Fig. 2. The final kinetic model for the compartmental analysis. Arrows represent pathways of fractional transport between the compartments. Delay is indicated with a jagged arrow.

Table 1. Model definition

Model equations (hidden stage)

$$X_S(t) = \frac{k_a k_e}{Cl(k_a - k_e)}\left(e^{-k_e t} - e^{-k_a t}\right)$$
$$\frac{d}{dt}X_I(t) = k_1 X_S(t) - k_3 X_I(t)$$
$$\frac{d}{dt}X_L(t) = k_3 X_I(t) - k_4 X_L(t)$$
$$\frac{d}{dt}X_P(t) = k_2 X_S(t) + k_4 X_L(t-\tau)$$

Model equations (observational stage)

$$S(t) = X_S(t) + 0.1X_L(t) + 0.1X_I(t) + 0.01X_P(t)$$
$$L(t) = a_L + 0.9(X_L(t) + 0.1X_P(t))$$
$$I(t) = a_I + 0.9(X_I(t) + 0.01X_P(t))$$
$$P(t) = X_P(t)$$

Observation equations

$$y_i^S = S(t_i) + (b^S + c^S S(t_i))\epsilon_i^S$$
$$y_i^I = I(t_i) + c^I I(t_i)\epsilon_i^I$$
$$y_i^L = L(t_i) + (b^L + c^L L(t_i))\epsilon_i^L$$
$$y_i^P = P(t_i) + c^P P(t_i)\epsilon_i^P$$
$$\epsilon_i^S, \epsilon_i^I, \epsilon_i^L, \epsilon_i^P \sim N(0,1)$$

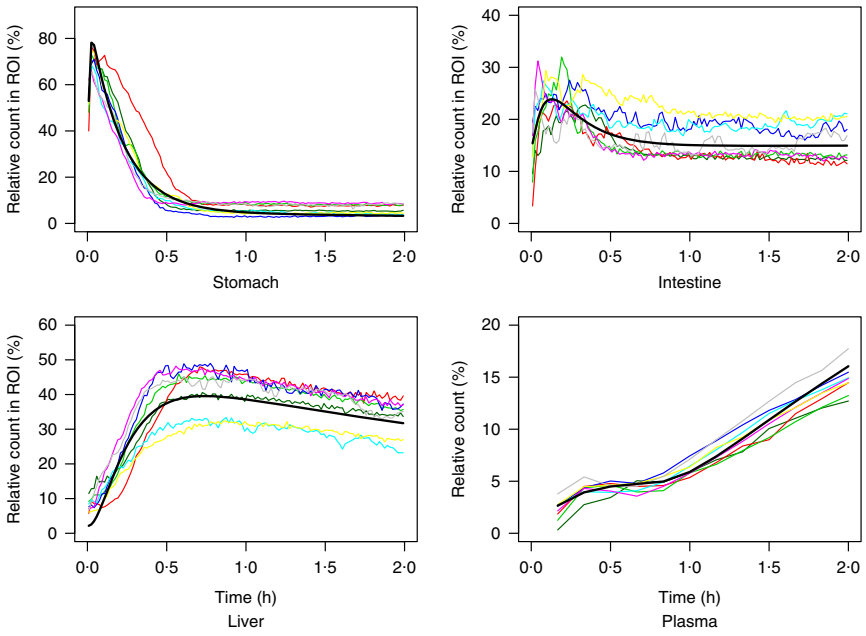


Fig. 3. Measured data for all participating subjects. ——, Subject A; ——, Subject B; ——, Subject C; ——, Subject D; ——, Subject E; ——, Subject F; ——, Subject G; ——, Subject H; ——, population fit. ROI, region of interest.

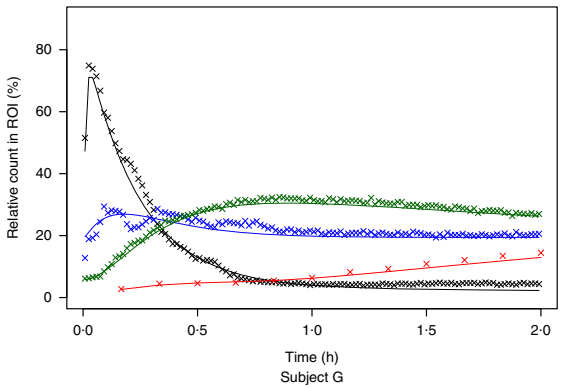


Fig. 4. Example of measured and modelled data for one participating subject. —×—, Stomach; —×—, intestine; —×—, liver; —×—, plasma. ROI, region of interest.

parameters along with their respective standard errors. Moreover, Table 2 depicts estimates of the measurement error parameters. The parameters indicated with *c* are proportions, and thus the largest measurement error was estimated to be in plasma, where it was 15·7 %. This is plausible, as the entire plasma count was extracted from a blood sample and prone to error due to estimation of blood volume and a possible heterogeneous distribution of SeMet in blood.

The mean disappearance half-life of ^{75}Se-SeMet in the liver was estimated to 2·65 (95 % CI 2·56, 2·76) h and the delay in the liver was 0·714 (95 % CI 0·640, 0·790) h.

Discussion

Although Se is recognised as a nutrient essential to human health, initial Se metabolism is poorly characterised. Currently, our understanding of Se absorption, whole-body retention and excretion is based on whole-body counting[5] or balance and tracer studies[6–8]. Data from previous studies[13–15] indicate a fundamental complexity of Se metabolism that can be explained by several factors. SeMet, the predominant form of Se in plant foods, is more easily absorbed compared with inorganic Se; however, both forms of Se are incorporated as selenocysteine into a variety of different selenoproteins, and Se is excreted in the urine in several forms.

Through compartmental analysis, it is possible to reduce the complexity of Se metabolism and to obtain an integrated picture of whole-body Se utilisation. On the basis of urine and faecal collection for 12 d and blood sampling for 120 d after oral ingestion of radio-labelled Se compounds, Wastney *et al.*[8] used compartmental modelling to provide new insight into human metabolism of Se: specifically, the number of metabolic pools and their sizes, relationships and turnover rates. To attain an acceptable fit of the raw data, they constructed a complex, multi-pool model. This may be consistent with the fact that the human selenoproteome contains at least twenty-five selenoproteins, which can be expected to have different turnover rates[14].

In this study, we focused on the intestinal absorption of ingested SeMet and its movements between a restricted number of pools over the subsequent 2 h. The overall strength of the study was that the compartmental analysis was based on data sampled with high temporal resolution directly from each tissue

Table 2. Definition of variables and parameters and estimated parameter values (Estimates with their standard errors)

		System variables		
$X_S(t)$, $S(t)$, y_i^S	Count	Tracer in the stomach (in tissue, in ROI, observed)		
$X_I(t)$, $I(t)$, y_i^I	Count	Tracer in the intestine (in tissue, in ROI, observed)		
$X_L(t)$, $L(t)$, y_i^L	Count	Tracer in the liver (in tissue, in ROI, observed)		
$X_P(t)$, $P(t)$, y_i^P	Count	Tracer in the plasma (in tissue, in tissue, observed)		
Parameters	Units	Explanation	Estimate	SE
		Population parameters		
a_I	Count	Residual level in the intestine	32·572	2·120
a_L	Count	Residual level in the liver	4·670	2·842
k_a	1/h	Absorption rate in the stomach	109·014	32·665
k_e	1/h	Total elimination rate from the stomach	4·894	0·225
Cl	1/h/count	Clearance/dose	0·025	0·002
k_1	1/h	Rate from the stomach to the intestine	2·630	0·291
k_2	1/h	Rate from the stomach to plasma	0·267	0·020
k_3	1/h	Rate from the intestine to the liver	13·199	2·313
k_4	1/h	Rate from the liver to plasma	0·261	0·005
τ	h	Delay in flow from the liver to plasma	0·714	0·037
		Parameters of observation error model		
b^S	Counts	Additive component in observation error for the stomach	5·042	0·187
c^S	1	Multiplicative component in observation error for the stomach	0·069	0·008
c^I	1	Multiplicative component in observation error for the intestine	0·084	0·002
b^L	Counts	Additive component in observation error for the liver	6·291	0·314
c^L	1	Multiplicative component in observation error for the liver	–0·036	0·003
c^P	1	Multiplicative component in observation error for plasma	0·157	0·012

ROI, region of interest.

of the model: three abdominal pools defined by the imaging technique and a plasma pool. We found that when restricting the investigation of kinetics to the first 2 h after SeMet administration, the dynamics of the data can be captured by a considerably simpler model as compared with the more complex models, which previously have been employed to describe Se metabolism[7,8]. The reason for this may be 2-fold. First, the measurements are taken on different time scales. Although measurements in the present study were taken every minute during the first 2 h after dose administration, the previous studies focused primarily on long-term dynamics, with data collection only at 30, 60 and 120 min during the first 2 h. These models, being designed for long-term behaviour, were therefore not able to describe the initial Se kinetics. Second, the previously applied doses of Se (150–200 μg) were considerably higher than those administered in the present study (29 μg). This difference in administered dose might result in a change of the kinetics.

According to our model, the initial amounts of radioactivity in the small intestine and liver compartments were on average 14·7 and 2·1 %, respectively, of the total dose of ^{75}Se-SeMet. These findings indicate the rapid flow of the first part of the tracer from the stomach to the small intestine and from the small intestine to the liver, which is additionally implied by the rate parameter estimates of k_1 and k_3. The parameter a_I was introduced to account for the final level of radioactivity remaining in the small intestine. This level was reached after about 60 min in all participants and comprised on average 14·7 % of the dose of ^{75}Se-SeMet and might reflect use of SeMet in the protein synthesis of the enterocytes. Lathrop *et al.*[16] determined the concentration of ^{75}Se-SeMet in the liver 2·4 h after ingestion to be about 13 % of the dose/kg, corresponding to a total of about 24 %. In agreement with this, our model predicted that, on average, 33 % of the total dose of ^{75}Se-SeMet was located in the liver 2 h after oral intake.

Consistent with previous observations[7], our raw data showed an almost monotone increase in the plasma concentration of ^{75}Se for the first 2 h after oral intake of ^{75}Se-SeMet in all participants. Between 20 and 40 min after ingestion, however, the concentration of ^{75}Se in the plasma reached a temporary plateau. This phenomenon, which has not been reported previously and most likely was exposed by the high temporal resolution of our data sampling, could actually be explained by our model. The model indicated that the first rise of Se level in plasma was due to inflow from the stomach, whereas the second rise was caused by inflow from the liver. The plateau originates from a delay in the flow from liver to plasma. Most likely, this delay could be explained by metabolic processes involving SeMet within the liver.

Note that the model proposed in the present analysis did not include any outflow from plasma. This assumption is certainly not in line with the true kinetics as the tracer will leave the body after a while. However, for the observed 2 h of study, this is in agreement with the findings of Swanson *et al.*[7] and Wastney *et al.*[8]. In their studies, the level of the tracer in plasma was consistently rising during the first couple of hours, and it did not start to decrease before approximately 3 h after administration.

The purpose of our study was to model a natural state in the kinetics of the underlying system. If the rates at which ^{75}Se-SeMet moves through the system are not constant, the findings reflect not only the rate at which ^{75}Se-SeMet itself moves from one compartment to another but also the changes in rate. To meet the requirements of a natural state, the dose of ^{75}Se-SeMet should be small relative to the amount of SeMet in

British Journal of Nutrition

the diet, so as not to change the natural metabolism in the system during the study. In the present study, all participating volunteers had normal plasma Se levels, and, apart from a short fast before data sampling, the study did not interfere with their usual diet regimen. Given that the normal daily dietary intake of Se in men is 60–120 μg[17], it is unlikely that our test dose of SeMet, which contained about 30 μg of Se, had a significant influence on the absorption and metabolism of SeMet *per se*. Consequently, it is reasonable to assume that all participants were in a natural state of SeMet turnover during data sampling.

In this study, fast dynamic imaging captured a relatively low count in each image. Inevitably, therefore, our procedure for defining the ROI was not perfect. Thus, a normal overlapping of parts of various organs in the anterior–posterior projections or movements of the participants during image acquisition could have caused some of the registered counts to be allocated to the incorrect ROI. However, we accounted for these deficiencies by explicitly including them in our final model.

The gamma camera technique is capable of tracing small quantities of γ ray emitting substances *in vivo*. Thus, the present dynamic approach provides an opportunity to explore the effects of food composition, gastrointestinal motility and gastrointestinal resection or bypass on the gastrointestinal absorption and initial turnover of physiological amounts of SeMet or other radio-labelled nutrients or food elements non-invasively. Thus, radio-labelled SeMet as a component of normal dietary protein could have yielded temporal information about the gastric emptying and the gastrointestinal breakdown of the dietary selenoproteins. Such data could have been incorporated into a kinetic model of the absorption and initial metabolism of dietary SeMet. However, the rate constants k_1 and k_3 derived using our imaging technique and modelling procedure may prove useful in clinical settings specifically focusing on the small intestine absorption capacity or aspects of the liver function.

Acknowledgements

The authors thank Ingelise Siegumfeldt and Bente Pedersen for their assistance in plasma analysis and all the subjects for their participation in the study. The mathematical processing was a part of the Dynamic Systems Interdisciplinary Network, University of Copenhagen.

The study was funded by grants from The Aase and Ejnar Danielsen Foundation (10-000243), The Hartmann Brothers Foundation (A7572), The P. A. Messerschmidt and Wife Foundation (028077-0002) and The Beckett Foundation (178PV/LS). None of the funders had any role in the design, analysis or writing of this article.

J. L. M. and M. D. designed the research; M. D., J. L. M., L. R. S. and S. F. conducted the research; M. G. R., S. D., L. R. S., S. F., J. T. O. and J. L. M. analysed the data; J. L. M., S. D. and M. G. R. wrote the paper; and J. L. M. had primary responsibility for the final content. All authors read and approved the final manuscript.

None of the authors declare any conflicts of interest.

References

1. Rayman MP (2000) The importance of selenium to human health. *Lancet* **356**, 233–241.
2. Papp LV, Lu J, Holmgren A, *et al.* (2007) From selenium to selenoproteins: synthesis, identity, and their role in human health. *Antioxid Redox Signal* **9**, 775–806.
3. Nickel A, Kottra G, Schmidt G, *et al.* (2009) Chacteristics of transport of selenoamino acids by epithelial amino acid transporters. *Chem Biol Interact* **177**, 234–241.
4. Fairweather-Tait SJ, Bao Y, Broadley MR, *et al.* (2011) Selenium in human health and disease. *Antioxid Redox Signal* **14**, 1337–1383.
5. Ben-Porath M, Case L & Kaplan E (1968) The biological half-life of ^{75}Se-selenomethionine in man. *J Nucl Med* **9**, 168–169.
6. Griffiths NM, Stewart RDH & Robinson MF (1976) The metabolism of [^{75}Se]selenomethionine in four women. *Br J Nutr* **35**, 373–382.
7. Swanson CA, Patterson BH, Levander OA, *et al.* (1991) Human [^{74}Se]selenomethionine metabolism: a kinetic model. *Am J Clin Nutr* **54**, 917–926.
8. Wastney ME, Combs GF Jr, Canfield WK, *et al.* (2011) A human model of selenium that integrates metabolism from selenite and selenomethionine. *J Nutr* **141**, 708–717.
9. Madsen JL, Sjögreen-Gleisner K, Elema DR, *et al.* (2014) Gamma camera imaging for studying intestinal absorption and whole-body absorption of selenomethionine. *Br J Nutr* **111**, 547–553.
10. Hurley PJ (1975) Red cell and plasma volumes in normal adults. *J Nucl Med* **16**, 46–52.
11. Lavielle M (2014) *Mixed Effects Models for the Population Approach: Models, Tasks, Methods and Tools*, (Chapman & Hall/CRC Biostatistics Series). London: Chapman and Hall/CRC.
12. Lixoft (2014) Monolix version 4.3.3. http://www.lixoft.eu
13. Burk RF, Norsworthy BK, Hill KE, *et al.* (2006) Effects of chemical form of selenium on plasma biomarkers in a high-dose human supplementation diet. *Cancer Epidemiol Biomarkers Prev* **15**, 804–810.
14. Kryukov GV, Castellano S, Novoselov SV, *et al.* (2003) Characterization of mammalian selenoproteomes. *Science* **300**, 1439–1443.
15. Francesconi KA & Pannier F (2004) Selenium metabolites in urine: a critical overview of past work and current status. *Clin Chem* **50**, 2240–2253.
16. Lathrop KA, Johnston RE, Blau M, *et al.* (1972) Radiation dose to humans from ^{75}Se-L-selenomethionine. *J Nucl Med* **6**, Suppl. 6, 10–26.
17. Flynn A, Hirvonen T, Mensink GBM, *et al.* (2009) Intake of selected nutrients from foods, from fortification and from supplements in various European countries. *Food Nutr Res* **53**, Suppl. I, 1–51.

British Journal of Nutrition

III - Inference for biomedical data using diffusion models with covariates and mixed effects

Submitted for publication.

Mareile Große Ruse
Department of Mathematical Sciences
University of Copenhagen, Copenhagen, Denmark

Susanne Ditlevsen
Department of Mathematical Sciences
University of Copenhagen, Copenhagen, Denmark

Adeline Samson
Laboratoire Jean Kuntzmann
Université Grenoble Alpes, Grenoble, France

Keywords: Approximate maximum likelihood, asymptotic normality, consistency, covariates, LAN, mixed effects, non-homogeneous observations, random effects, stochastic differential equations, EEG

Inference for biomedical data using diffusion models with covariates and mixed effects

Mareile Große Ruse

Department of Mathematical Sciences, University of Copenhagen.
mareile@math.ku.dk

Adeline Samson

Univ. Grenoble Alpes, LJK, F-38000 Grenoble, France.
adeline.leclercq-samson@imag.fr

Susanne Ditlevsen

Department of Mathematical Sciences, University of Copenhagen.
susanne@math.ku.dk

Summary. Neurobiological data such as EEG measurements pose a statistical challenge due to low spatial resolution and poor signal-to-noise ratio, as well as large variability from subject to subject. We propose a new modeling framework for this type of data based on stochastic processes. Stochastic differential equations with mixed effects are a popular framework for modeling biomedical data, e.g., in pharmacological studies. While the inherent stochasticity of diffusion models accounts for prevalent model uncertainty or misspecification, random effects models inter-subject variability. The 2-layer stochasticity, however, renders parameter inference challenging. This is especially true for more complex model dynamics, and only few theoretical investigations on the asymptotic behavior of estimates exist. This article adds to filling this gap by examining asymptotics (number of subjects going to infinity) of Maximum Likelihood estimators in multidimensional, nonlinear and non-homogeneous stochastic differential equations with random effects and included covariates. Estimates are based on the discretized continuous-time likelihood and we investigate finite-sample and discretization bias. In applications, the comparison of, e.g., treatment effects, is often of interest. We discuss hypothesis testing and evaluate by simulations. Finally, we apply the framework to a statistical investigation of EEG recordings from epileptic patients.

Keywords: Approximate maximum likelihood, asymptotic normality, consistency, covariates, LAN, mixed effects, non-homogeneous observations, random effects, stochastic differential equations, EEG

1. Introduction

Many biomedical studies are based on image data, which is characterized by a high time resolution, but also a low signal-to-noise ratio. The same happens with EEG data, which are measurements of electrical activitity measured from electrodes on the scalp, and are proxies of underlying brain activity. This high frequency and noisy nature of the data lends itself naturally to be modeled by continous-time stochastic processes. Moreover, data are often multi-dimensional and repeated on a collection of subjects. The noise may

be due to factors such as internal and external fluctuations, difficult experimental conditions, or a collection of multiple unmeasured effects, for example non-specified feedback mechanisms or genetic variation. The intra-subject variability in longitudinal data asks for a model that incorporates system noise. Any systematic inter-subject variability is usually well explained by the inclusion of covariate information, e.g., treatment regime, gender or specifics of experimental conditions. The remaining inter-subject variability can then be taken care of by random effects. The motivating examples for our work are EEG measurements from multiple channels, and a compartment model arising in a recent pharmacological study based on image data. Both types of data are measured at high frequency, i.e., the sampling frequency is fast compared to the typical time scales of the observed system. This allows us to employ techniques facilitating the use of continuous-time stochastic processes. We therefore propose a new modeling framework where the observed time series are assumed to be generated from a multi-dimensional stochastic differential equation (SDE), which accounts for systematic and random inter-subject variability through covariates and random effects.

Models that combine SDEs and random effects (i.e., so-called stochastic differential mixed-effects models, SDMEMs) have become a popular framework for modeling biological data (Guy et al., 2015; Donnet et al., 2010; Møller et al., 2010; Leander et al., 2014; Picchini et al., 2008). They come with three advantages: Firstly, they capture inter-subject variations by incorporation of random effects. Secondly, they account for model uncertainty or environmental fluctuations by their inherent stochasticity. Lastly, they remedy the otherwise omnipresent issue of the inconsistent drift estimator (Kessler et al., 2012) in plain SDEs (only fixed effects), when the observation time horizon is finite. The latter is due to the fact that the mixed-effects approach facilitates pooling of data across subjects, which leads to unbiasedness of the drift estimator as the number of subjects approaches infinity.

However, the flexibility and robustness of SDMEMs come at a price and bear particular challenges in terms of statistical inference. The data likelihood in these models is generally intractable, for two reasons: On the one hand, the likelihood for (nonlinear) SDE models is analytically not available, rendering parameter inference for standard SDE models a nontrivial problem in itself. On the other hand, the likelihood has to be integrated over the distribution of the random effects. Thus, numerical or analytical approximations are inevitable. The likelihood for SDE models can be approximated in various ways. Given discrete-time observations, the likelihood is expressed in terms of the transition density. Approximation methods for the latter reach from solving the Fokker-Planck equation numerically (Lo, 1988), over standard first-order (Euler-Maruyama) or higher-order approximation schemes and simulation-based approaches (Pedersen, 1995; Durham and Gallant, 2002) to a closed-form approximation via Hermite polynomial expansion (Aït-Sahalia, 2002). If continuous-time observations are assumed (e.g., if high-frequency data is available), transition densities are not needed and the likelihood can be obtained from the Girsanov formula (Phillips and Yu, 2009). Popular analytical approximation techniques for general nonlinear mixed-effects models are first-order conditional estimation (FOCE) (Beal and Sheiner, 1981) and Laplace approximation (Wolfinger, 1993). A computational alternative is the expectation-maximization (EM) algorithm, or stochastic versions thereof (Delyon et al., 1999).

In the context of SDMEMs, the above mentioned approximation methods have been combined in various ways, depending on whether observations are modeled in discrete or in continuous time (here we do not consider measurement noise). For discrete-time observations, Hermite expansion of the transition density has been combined with Gaussian quadrature algorithms and Laplace's approximation (Picchini et al., 2010; Picchini and Ditlevsen, 2011). Mixed effects that enter the diffusion coefficient are investigated in Delattre et al. (2015, 2017). The case of continuous-time observations of a univariate SDMEM with Gaussian mixed effects entering the drift linearly is considered in Delattre et al. (2013).

Two aspects that are important in modeling biomedical data are not covered by these works: On the one hand, the theoretical investigations of estimators when the state process is modeled by a multivariate, time-inhomogeneous and nonlinear SDE, and on the other hand, the inclusion of covariate information. The lack of both in a model implies considerable restrictions for practitioners and the purpose of this article is to fill this gap.

If the drift function is linear in the parameters, the standard asymptotic properties of the MLE in multidimensional, time-homogeneous, nonlinear SDMEMs can be shown by a natural extension of the proofs in Delattre et al. (2013). In particular, the model likelihood turns into a neat expression, and all remaining model complexities (multidimensionality of the state, nonlinearity, covariates) are conveniently hidden in the sufficient statistics. The results in Delattre et al. (2013) on the discretization error which arises when continuous-time statistics are replaced by their discrete-time versions hold as well in the more complex model setup. Their approach has, however, two drawbacks. The first one is model-related: It is assumed that observations are identically distributed, which impairs the inclusion of subject-specific covariate information. The other drawback is proof-related: The imposed regularity assumptions are rather restrictive, for instance, the density of the random effects may not be smooth. If, for instance, random effects are supposed to have a double exponential distribution, those regularity conditions are not met. However, the Laplace density, e.g., is "almost" regular, satisfying a particular type of first-order differentiability. This almost regularity can be treated by the more general approach which builds upon L_2-differentiability and the local asymptotic normality (LAN) property of a sequence of statistical models (Le Cam, 2012; Ibragimov and Has'minskii, 2013). Therefore, we approach the theoretical investigations from the more general LAN perspective.

In regression models, the convergence of the average Fisher information is a standard assumption which facilitates the verification of MLE asymptotics considerably. We address this condition in the SDMEM setup and point out the difficulties that arise here, when observations are not identically distributed.

The article is structured as follows. Section 2 introduces the model framework, investigates asymptotic properties of the MLE and applies the asymptotic results to hypothesis testing. Moreover, we exemplify the framework with covariates for affine mixed effects. Section 3 is devoted to a simulation study in a model which is common in pharmacokinetics and is motivated by a recent study on Selenium metabolism in humans (Große Ruse et al., 2015). Here, we study finite sample and discretization bias of the estimation procedure and properties regarding hypothesis testing, where we investigate the effect

of a drug treatment (as encoded by a covariate with levels *treatment* and *placebo*). We then apply the SDMEM framework to EEG recordings of epileptic patients in section 4, with the purpose of investigating how channel interactions differ between non-seizure and seizure states. Finally, we conclude with a discussion.

2. Maximum likelihood estimation for SDMEMs with covariates

This section considers parameter inference when observations are independent, but not necessarily identically distributed, a setting that naturally occurs when covariate information is included in the model formulation.

2.1. Model formulation

We consider N r-dimensional stochastic processes $X^i = (X_t^i)_{0\le t\le T^i}$ whose dynamics are governed by the stochastic differential equations

$$dX_t^i = F(X_t^i, D_t^i, \mu, \phi^i)dt + \Sigma(t, X_t^i)dW_t^i, \quad 0 \le t \le T^i, \quad X_0^i = x_0^i, \quad i = 1, \dots, N. \quad (1)$$

The r-dimensional Wiener processes $W^i = (W_t^i)_{t\ge 0}$ and the d-dimensional random vectors ϕ^i are defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\ge 0}, \mathbb{P})$, which is rich enough to ensure independence of all random objects $W^i, \phi^i, i = 1, \dots, N$. The d-dimensional vectors $\phi^i, i = 1, \dots, N$, are the so-called *random effects*. They are assumed to be $\mathcal{F}_0$-measurable and have a common (usually centered) distribution which is specified by a (parametrized) Lebesgue density $g(\varphi; \vartheta)d\varphi$. The parameter $\vartheta \in \mathbb{R}^{q-p}$ is unknown, as well as the *fixed effect* $\mu \in \mathbb{R}^p$. The combined parameter $\theta = (\mu, \vartheta)$ is the object of statistical inference and is assumed to lie in the parameter space Θ, which is a bounded subset of $\mathbb{R}^q$. The $D^i : [0, T^i] \to \mathbb{R}^s$ encode subject-specific covariate information and are assumed to be known. They can also encode a general time dependency, which not necessarily is subject specific. The functions $F : \mathbb{R}^{r+s+p+d} \to \mathbb{R}^r, \Sigma : [0, T] \times \mathbb{R}^r \to \mathbb{R}^{r\times r}$, with $T = \max_{1\le i\le N} T^i$, are deterministic and known and the initial conditions x_0^i are r-dimensional random vectors. We assume standard regularity assumptions on the drift (including the D^i) and diffusion functions to assure (i) existence and uniqueness of the solution to (1) and (ii) existence and good behaviour of the Radon-Nikodym derivative

$$\begin{aligned} q^i(\mu, \varphi) &:= q^i(\mu, \varphi; X^i) = \frac{d\mathbb{Q}^i_{\mu,\varphi}}{d\mathbb{Q}^i_{\mu_0,\varphi_0}}(X^i) \\ &= \exp\left(\int_0^{T^i} \left[F(X_s^i, D_s^i, \mu, \varphi) - F(X_s^i, D_s^i, \mu_0, \varphi_0)\right]' \Gamma^{-1}(s, X_s^i) dX_s^i \right. \\ &\left. - \frac{1}{2}\int_0^{T^i} \left[F(X_s^i, D_s^i, \mu, \varphi) - F(X_s^i, D_s^i, \mu_0, \varphi_0)\right]' \Gamma^{-1}(s, X_s^i) \left[F(X_s^i, D_s^i, \mu, \varphi) + F(X_s^i, D_s^i, \mu_0, \varphi_0)\right] ds \right), \end{aligned}$$

where $\Gamma = \Sigma\Sigma'$ and $\mathbb{Q}^i_{\mu,\varphi}$ is the distribution of X^i conditioned on an observed $\phi^i = \varphi$ (and μ_0, φ_0 are fixed). The function q^i is the *conditional likelihood* for subject i given we have observed the random effect $\phi^i = \varphi$. Therefore, the *unconditional likelihood* for subject i is $p^i(\theta) := p^i(\theta; X^i) = \int_{\mathbb{R}^d} q^i(\mu, \varphi) \cdot g(\varphi; \vartheta)\, d\varphi$.

We observe X^i at time points $0 \le t_0^i < t_1^i < \dots < t_{n_i}^i = T^i$ and the inference task consists in recovering the "true" underlying θ based on observations $X_{t_0^i}^i, \dots, X_{t_n^i}^i, i =$

$1, \ldots, N$. We approach this inference task by first supposing to have the entire paths $(X_t^i)_{0 \le t \le T^i}, i = 1, \ldots, N$, at our disposal. Based on these we derive the continuous-time MLE and discretize it in a second step. The bias introduced by the discretization is investigated theoretically and by simulations.

2.2. Affine Gaussian mixed effects

In many applications the fixed and random effects enter the drift in an affine manner,

$$F(X_t^i, D_t^i, \mu, \phi^i) = A(X_t^i, D_t^i) + B(X_t^i, D_t^i)\,\mu + C(X_t^i, D_t^i)\,\phi^i. \tag{2}$$

An example of (2) is a widely used class of compartment models, which we illustrate in a simulation study in Section 3, and in our main application in section 4, where we analyze EEG data from epileptic patients. Likelihood-based inference then becomes explicit if the random effects are Gaussian distributed, $g(\varphi; \Omega) = \mathcal{N}(0, \Omega)(\varphi)$. The separation of μ and ϕ^i in (2) enables the modeler to impose random effects on only a selection of fixed effects. The conditional likelihood turns into the compact expression $q^i(\mu, \varphi) = \mathrm{e}^{\mu' U_{1i} - \frac{1}{2}\mu' V_{1i}\mu + \varphi' U_{2i} - \frac{1}{2}\varphi' V_{2i}\varphi - \varphi' Z_i \mu}$ with the sufficient statistics

$$\begin{aligned}
U_{1i} &= \int_0^{T^i} B(X_s^i, D_s^i)' \Gamma^{-1}(s, X_s^i) \left[dX_s^i - A(X_s^i, D_s^i) ds \right], \\
V_{1i} &= \int_0^{T^i} B(X_s^i, D_s^i)' \Gamma^{-1}(s, X_s^i) B(X_s^i, D_s^i) ds, \\
U_{2i} &= \int_0^{T^i} C(X_s^i, D_s^i)' \Gamma^{-1}(s, X_s^i) \left[dX_s^i - A(X_s^i, D_s^i) ds \right], \\
V_{2i} &= \int_0^{T^i} C(X_s^i, D_s^i)' \Gamma^{-1}(s, X_s^i) C(X_s^i, D_s^i) ds, \\
Z_i &= \int_0^{T^i} C(X_s^i, D_s^i)' \Gamma^{-1}(s, X_s^i) B(X_s^i, D_s^i) ds.
\end{aligned} \tag{3}$$

Integration over φ gives the unconditional likelihood for subject i,

$$p^i(\theta) = \frac{1}{\sqrt{\det(I + V_{2i}\Omega)}} \exp\left(\left[U_{1i}' - U_{2i}' R^i(\Omega) Z_i \right] \mu - \frac{1}{2}\mu' \left[V_{1i} - Z_i' R^i(\Omega) Z_i \right] \mu + \frac{1}{2} U_{2i}' R^i(\Omega) U_{2i} \right), \tag{4}$$

with $R^i(\Omega) = (V_{2i} + \Omega^{-1})^{-1}$. In particular, the MLE $\hat{\mu}_N$ of the fixed effect (given Ω) is explicit,

$$\hat{\mu}_N(\Omega) = \left[\sum_{i=1}^N \left[V_{1i} - Z_i' R^i(\Omega) Z_i \right] \right]^{-1} \left[\sum_{i=1}^N \left[U_{1i} - Z_i' R^i(\Omega) U_{2i} \right] \right]. \tag{5}$$

REMARK 1. *The likelihood p^i is explicit even if the fixed effect enters the drift nonlinearly. However, only a linear fixed effect μ leads to an explicit expression for its MLE.*

Discrete data

Above we assumed to observe the entire paths $(X_t^i)_{0 \le t \le T}$. In practice, observations are only available at discrete time points $t_0, \ldots, t_n$. A natural approach is to replace the continuous-time

integrals in $q^i(\theta)$ by discrete-time approximations and to derive an approximate MLE based on the resulting approximate likelihood. For instance, an expression of the form $\int_{t_k}^{t_{k+1}} h(s, X_s^i) dX_s^i$ may be replaced by a first-order approximation $h(t_k, X_k^i)\Delta X_k^i$. In the linear model (2), the approximation of the continuous-time likelihood corresponds to the exact likelihood of its Euler scheme approximation. In particular, if we observe all individuals at time points $t_k = T\frac{k}{n}$ and denote by $U_{1i}^n, V_{1i}^n, U_{2i}^n, V_{2i}^n, Z_i^n$ the first-order discrete-time approximations to the continuous-time statistics $U_{1i}, V_{1i}, U_{2i}, V_{2i}, Z_i$ in eq. (3), one has the following result:

THEOREM 1 (NEGLIGIBILITY OF DISCRETIZATION ERROR).
Assume model (2) *and suppose that* $A, B'\Gamma^{-1}B, B'\Gamma^{-1}C, C'\Gamma^{-1}C, B'\Gamma^{-1}, C'\Gamma^{-1}$ *are globally Lipschitz-continuous in* (t, x) *and that in addition to* A, B, C *and* Σ *also* $B'\Gamma^{-1}, C'\Gamma^{-1}$ *is of sublinear growth in* x*, uniformly in* t*. Then, for all* $p \geq 1$ *and all* $i = 1, \ldots, N$*, there is a constant* K *such that*

$$\mathbb{E}_{\theta_0}\left([\![V_{1i} - V_{1i}^n]\!]^p + \|U_{1i} - U_{1i}^n\|^p + [\![V_{2i} - V_{2i}^n]\!]^p + \|U_{2i} - U_{2i}^n\|^p + \|Z_i - Z_i^n\|^p\right) \leq K\left(\frac{T}{n}\right)^{p/2}.$$

The discretization error is investigated numerically in section 3.

2.3. Asymptotic properties of the MLE

If the drift is as in (2) and observations are identically distributed (in particular, the model does not contain subject-specific covariate information), consistency and asymptotic normality of the MLE can be proved using the ideas in Delattre et al. (2013). The proofs are a natural extension of their setting to the multidimensional, affine, non-homogeneous case, but become more tedious to work out in detail and to write down and will therefore be omitted here. We will get back to the affine model with i.i.d. observations in subsection 2.3.1 and in section 3.

The classical proof of asymptotic normality of the MLE imposes strong smoothness conditions on the subject-specific density functions, such as third-order differentiability and boundedness of the derivatives. A Taylor expansion argument together with a required asymptotic normality of the N-sample Score function and a convergence of the average Fisher Information (FI) (see, e.g., Bradley and Gart (1962, equation (13)), or Hoadley (1971, condition N7)) then yield the result. If observations are not identically distributed (e.g., if subject-specific covariate information is included in (1)) and the standard central limit theorem for i.i.d. variables can not be applied to the Score function, one can revert to the Lindeberg-Feller central limit theorem, given the family of individual score functions $\{S^i(\theta); i \in \mathbb{N}\}$ satisfies the Lindeberg condition (a condition which limits the variation of each S^i in relation to the overall N-sample score variation). The convergence of the average FI, which is naturally given in i.i.d. models, often breaks down to requiring that covariate averages converge (Fahrmeir and Kaufmann, 1985).

The more general LAN approach which we pursue here dispenses with the differentiability conditions by building upon L_2-derivatives. An L_2-Score function and L_2-Fisher information are defined, which then are required to meet the above mentioned Lindeberg and convergence conditions (cf. assumption (e) below and Theorem 3). The first part of this section adapts results developed in Ibragimov and Has'minskii (2013), on consistency and asymptotic normality of the MLE for $\theta = (\mu, \vartheta)$ in models that do not necessarily meet the differentiability conditions, to the current framework of SDMEMs with covariates. In a second part, we illustrate the verification of regularity conditions for an SDMEM with covariates and with dynamics that are frequently encountered in biomedical modeling. While the L_2-based approach opens up for the inclusion of irregular densities into our framework, it still requires one to verify the convergence of the

average FI. We will discuss the complications of the latter within the SDMEM framework at the end of this section.

We write $\nu^i = \mathbb{Q}^i_{\mu_0,\varphi_0}$ (see beginning of section 2). For simplicity, we assume that $\Theta \subseteq \mathbb{R}^q$ is open, bounded and convex and that in all what follows, $K \subset \Theta$ is compact.

We start by stating general assumptions which the statistical model is required to satisfy and adapt them more closely to the SDMEM framework, by pointing out sufficient conditions for this particular framework which may be verified more easily. Afterwards, we establish results on asymptotic properties of the MLE for SDMEMs.

(a) $\theta \mapsto p^i(\theta)$ is ν^i-a.s. continuous.
(b) $\theta \mapsto \sqrt{p^i(\theta)}$ is $L_2(\nu^i)$-differentiable† with $L_2(\nu^i)$-derivative $\psi^i(\theta)$ (in other words: $p^i(\theta)$ is Hellinger differentiable).
(c) $\psi^i(\theta)$ is continuous in $L_2(\nu^i)$.
As a consequence, the matrix $I^i(\theta) = 4\int \psi^i(\theta;x)'\psi^i(\theta;x)d\nu^i(x)$ exists and is continuous and will be called the FI matrix. The N-sample FI is then $I_N(\theta) = \sum_{i=1}^N I^i(\theta)$.
(d) The FI is bounded away from 0 and finite: $0 < \inf_{\theta\in\Theta}[\![\frac{1}{N} I_N(\theta)]\!] \leq \sup_{\theta\in\Theta}[\![\frac{1}{N} I_N(\theta)]\!] < \infty$.
(e) There is a symmetric, positive definite limiting matrix $I(\theta)$ such that
$\lim_{N\to\infty} \sup_{\theta\in K}[\![\frac{1}{N} I_N(\theta) - I(\theta)]\!] = 0$ and $\lim_{N\to\infty} \sup_{\theta\in K}[\![\big(\frac{1}{N} I_N(\theta)\big)^{-1/2} - I(\theta)^{-1/2}]\!] = 0$.

Analogously to the traditional setting, we call $S^i(\theta) = 2p^i(\theta)^{-1/2}\psi^i(\theta)$ the score function of sample i and set $S_N(\theta) = \sum_{i=1}^N S^i(\theta)$ for the N-sample score function. One can show that also in this more general setting the score function is centered (Ibragimov and Has'minskii, 2013, p. 115).

Sufficient conditions for the a.s. continuity of $p^i(\theta)$ in θ are continuity of $\mu \mapsto q^i(\mu,\varphi)$ and of $\vartheta \mapsto g(\varphi;\vartheta)$, together with the existence of an integrable function of φ dominating $q^i(\mu,\varphi)g(\varphi;\vartheta)$. Continuity of g holds for instance in the common case where g is a Gaussian density $\mathcal{N}(0,\vartheta)$ and ϑ is bounded away from 0. For conditions on the continuity of q^i, suppose F is continuous and assume for simplicity $\Sigma(t,x) \equiv I$ is the identity matrix. If $\mu \mapsto F(X^i_s, D^i_s, \mu, \varphi)$ is uniformly continuous (for instance differentiable with bounded Jacobian), then $\mu \mapsto \int_0^{T^i} F(X^i_s, D^i_s, \mu, \varphi)'F(X^i_s, D^i_s, \mu, \varphi)ds$ is continuous. If F moreover has the property $\big\|F(X^i, D^i_s, \mu, \varphi) - F(X^i, D^i_s, \mu_0, \varphi)\big\| \leq K(1 + \big\|X^i\big\|^\kappa)\,\|\mu - \mu_0\|$ for some $\kappa > 0$, Kolmogorov's continuity criterion guarantees continuity of q^i.

The L_2-differentiability is neither stronger nor weaker than standard (point-wise) differentiability. Generally, none implies the other, but under certain conditions, the limits are identical. Of course, if p^i is L_2-differentiable and differentiable in the ordinary sense, then $\psi^i(\theta;x) = \frac{d}{d\theta}\left[p^i(\theta;x)^{1/2}\right]$.

To point out the connection between the FI and score functions defined via L_2-derivatives and their counterparts based on "standard" differentiability, we recall the following result (Van der Vaart, 2000, Lemma 7.6). If $\theta \mapsto \sqrt{p^i(\theta)}$ is continuously differentiable, the quantity $\tilde{S}^i(\theta) := 2p^i(\theta)^{-1/2}\frac{d}{d\theta}p^i(\theta)$ is well-defined (since $p^i > 0$). If $\tilde{I}^i(\theta) = \mathbb{E}_\theta(\tilde{S}^i(\theta)\tilde{S}^i(\theta)')$ is finite and continuous, $\theta \mapsto \sqrt{p^i(\theta)}$ is L_2-differentiable, the L_2-derivative coincides with the point-wise derivative and in fact, $\tilde{S}^i(\theta) = S^i(\theta)$ and $\tilde{I}^i(\theta) = I^i(\theta)$.

Note as well that the assumption on the (norm of the) Fisher information matrix to grow beyond bounds (cf. condition (e)) corresponds to the requirement of infinite flow of information. This is naturally connected to the consistency of estimators.

†For each θ, $\int\|\psi^i(\theta;x)\|^2 d\nu^i(x) < \infty$ and $\lim_{\|h\|\to 0}\|h\|^{-2}\int\|\sqrt{p^i(\theta+h;x)} - \sqrt{p^i(\theta;x)} - \psi^i(\theta;x)h\|^2 d\nu^i(x) = 0$.

In the sequel, we write shortly and somewhat sloppily θ_N if it is of the form $\theta_N = \theta + I_N(\theta)^{-1/2}h$ for some $\theta \in K$ and $h \in \Theta_{N,\theta} = \{h \in \mathbb{R}^q : \theta + I_N(\theta)^{-1/2}h \in \Theta\}$.

We are now in the position to state results on the asymptotic behavior of the MLE in SDMEMs with covariates. These are consequences of theorems in Ibragimov and Has'minskii (2013), and proofs are only shortly outlined.

THEOREM 2 (CONSISTENCY). *The MLE of model* (1) *is uniformly on* K *consistent, if*

(A.1) There is a constant $m > q$ *such that* $\sup_{\theta\in K} \mathbb{E}_\theta \left(\|S_N(\theta)\|^m\right) < \infty$.

(A.2) There is a positive constant $a(K)$ *such that for (sufficiently large* N *and) all* $\theta \in K$ *(and all* $h \in \Theta_{N,\theta}$*)* $H_i^2(\theta, \theta_N) \geq a(K)\frac{\|\theta_N-\theta\|^2}{1+\|\theta_N-\theta\|^2}$*, where* $H_i^2(\theta_1, \theta_2) := \int \left(\sqrt{p^i(\theta_1)} - \sqrt{p^i(\theta_2)}\right)^2 d\nu^i$ *is the squared Hellinger distance between* $\mathbb{Q}^i_{\theta_1}$ *and* $\mathbb{Q}^i_{\theta_2}$.

PROOF. (A.1) is an extension of Lemma III.3.2. in Ibragimov and Has'minskii (2013) to non-homogeneous observations. (A.2) is adapted from (Ibragimov and Has'minskii, 2013, Lemma I.5.3).

REMARK 2. *If the dimension of the parameter set is 1, (A.1) can be replaced by a sub-quadratic growth condition on the Hellinger distance (for i.i.d. observations, see Ibragimov and Has'minskii (2013, Theorem I.5.3)), namely that* $H^2(\theta_1, \theta_2) \leq A\|\theta_2-\theta_1\|^2$*, such that consistency here reduces to* $H^2(\theta_1, \theta_2)$ *behaving asymptotically as* $\|\theta_2 - \theta_1\|^2$.

The following theorem establishes the so-called uniform asymptotic normality of the model, which in turn implies the asymptotic normality of the MLE (Thms II.6.2. and III.1.1, Ibragimov and Has'minskii (2013)).

THEOREM 3 (ASYMPTOTIC NORMALITY). *Assume (A.1) and (A.2) from Theorem 2 and additionally*

(B.1) The family $\{S^i(\theta), i = 1, \ldots, N\}$ *satisfies the Lyapunov condition uniformly in* K*, i.e., there is* $\delta > 0$ *such that* $\lim_{N\to\infty} \sup_{\theta\in K} \sum_{i=1}^N \mathbb{E}_\theta \left(\left\|I_N(\theta)^{-1/2}S^i(\theta)\right\|^{2+\delta}\right) = 0$.

(B.2) $\forall R > 0 : \lim_{N\to\infty} \sup_{\theta\in K} \sup_{\|h\|<R} \sum_{i=1}^N \int \left(\left[\psi^i(\theta_N) - \psi^i(\theta)\right] I_N(\theta)^{-1/2}h\right)^2 d\nu^i = 0$.

Then $\{\hat\theta_N\}_{N\in\mathbb{N}}$ *is uniformly in* K *consistent, asymptotically Gaussian distributed with parameters* $(\theta, I_N(\theta)^{-1})$ *and all moments of* $\{I_N(\theta)^{1/2}(\hat\theta_N - \theta)\}_{N\in\mathbb{N}}$ *converge uniformly in* K *to the corresponding moments of the* $\mathcal{N}(0, I)$ *distribution.*

Condition (B.1) can be generalized to the Lindeberg condition. If the densities $\sqrt{p^i(\theta)}$ are twice continuously differentiable with second derivative $J^i(\theta)$, (B.2) can be replaced by requiring that $\lim_{N\to\infty} \sup_{\theta\in K} \sup_{\|h\|\leq R} [\![I_N(\theta)^{-1/2}]\!]^4 \sum_{i=1}^N \int [\![J^i(\theta_N)]\!]^2 d\nu^i = 0$. As pointed out in the introduction, for a general SDMEM the p^i are not explicitly available. One can, however, formulate conditions for the drift function F and the random effects density g, which implicitly guarantee the differentiability of $\log p^i(\theta) = \log\left(\int q^i(\mu, \varphi)g(\varphi; \vartheta)d\varphi\right)$. This can, for example, be done by assuring that differentiation can be passed under the integral sign: Sufficient conditions for the differentiability of $\log p^i(\theta)$ with respect to μ would, e.g., include differentiability of q^i w.r.t. μ and a uniform in μ domination of $\frac{d}{d\mu}q^i(\mu, \varphi)(\int q^i(\mu, \varphi)g(\varphi; \vartheta)d\varphi)^{-1}$. However, explicitly formulating these conditions for a generic SDMEM is not illustrative, suitable conditions should be formulated and checked for the specific application at hand. One particular case in which the $p^i(\theta)$ are explicitly available is the affine model (2), which we consider in more detail below.

2.3.1. SDMEM with covariates and affine mixed effects

We illustrate the verification of the assumptions for Theorems 2 and 3 for the affine SDMEM (2). This model, which certainly is more regular than required, will be revisited in section 3, where we study estimation performance and hypothesis testing for different sample sizes and sampling frequencies, and will be revisited in section 4 for the statistical investigation of EEG data. For simplicity we assume $B = C$, such that $U_i := U_{1i} = U_{2i}$ and $V_i = V_{1i} = V_{2i} = Z_i$. The likelihood (4) can be written as

$$p^i(\theta) = \frac{1}{\sqrt{\det(I + V_i\Omega)}} \exp\left(-\frac{1}{2}(\mu - V_i^{-1}U_i)' G^i(\Omega)(\mu - V_i^{-1}U_i)\right) \exp\left(\frac{1}{2}U_i' V_i^{-1} U_i\right).$$

with $G^i(\Omega) = (I + V_i\Omega)^{-1}V_i$. Defining $\gamma^i(\theta) = G^i(\Omega)(V_i^{-1}U_i - \mu)$ (we assume that V_i is a.s. invertible), the score function for subject i is thus given by $S^i(\theta) = \left[\frac{d}{d\mu}\log p^i(\theta), \frac{d}{d\Omega}\log p^i(\Omega)'\right]$, with

$$\frac{d}{d\mu}\log p^i(\theta) = \gamma^i(\theta)' \qquad \text{and} \qquad \frac{d}{d\Omega}\log p^i(\theta) = \frac{1}{2}\left[-G^i(\Omega) + \gamma^i(\theta)\gamma^i(\theta)'\right].$$

We start by verifying condition (A.1). Since the set $K \subset \Theta$ is compact, there are positive constants A_K, B_K, C_K such that $\|\mu\| \le A_K, B_K \le [\![\Omega]\!] \le C_K$. One can show that $[\![G^i(\Omega)]\!] \le [\![\Omega^{-1}]\!]$, which gives the upper bound $\|S^i(\theta)\| \le \left(\|\gamma^i(\theta)\| + [\![\Omega^{-1}]\!] + \|\gamma^i(\theta)\|^2\right)$. Moreover, the moment-generating function $\Phi_{\theta,\gamma^i(\theta)}(a)$ of $\gamma^i(\theta)$ can be bounded from above by $\mathrm{e}^{\frac{1}{2}a'\Omega^{-1}a}$, for $a \in \mathbb{R}^d$. This can be used to find that $\mathbb{E}_\theta\left(\|\gamma_i(\theta)\|^m\right) \le C_1$ for some constant C_1 that may depend on K, d, m. Therefore, there is another constant C_2, which may depend on K, d, m, N, such that $\mathbb{E}_\theta\left(\|S_N(\theta)\|^m\right) \le C_2$, proving (A.1).

To verify (A.2), note that the regularity of $p_N(\theta) = \prod_{i=1}^N p^i(\theta)$ and its derivatives implies that

$$\begin{aligned}
H^2(\theta, \theta_N) &= \int \left[-\psi_N(\theta)(\theta_N - \theta) + \left(\sqrt{p_N(\theta_N)} - \sqrt{p_N(\theta)}\right) + \psi_N(\theta)(\theta_N - \theta)\right]^2 d\nu \\
&= \int \left[-\psi_N(\theta)(\theta_N - \theta)\right]^2 d\nu + o(\|\theta_N - \theta\|^2) \\
&= (\theta_N - \theta)' I_N(\theta)(\theta_N - \theta) + o(\|\theta_N - \theta\|^2) - 2O(\|\theta_N - \theta\|^2)o(\|\theta_N - \theta\|^2) \\
&\ge \|(\theta_N - \theta)\|^2 \lambda_{N,\min}(\theta) + o(\|\theta_N - \theta\|^2).
\end{aligned}$$

where $\lambda_{N,\min}(\theta)$ denotes the smallest eigenvalue of $I_N(\theta)$. Therefore, for N sufficiently large, there is a constant A_K such that $H^2(\theta, \theta_N) \ge A_K\|(\theta_N - \theta)\|^2$. Since Θ is bounded, we even have $\|(\theta_N - \theta)\|^2 \ge C\frac{\|(\theta_N - \theta)\|^2}{1 + \|(\theta_N - \theta)\|^2}$ for some positive constant C, which shows that (A.2) holds.

The Lyapunov condition (B.1) follows in a straightforward way. According to the above, $\mathbb{E}_\theta\left(\|S^i(\theta)\|^3\right) \le C$ for some C and therefore

$$\begin{aligned}
\sup_{\theta\in K}\sum_{i=1}^N \mathbb{E}_\theta\left(\|I_N(\theta)^{-1/2}S^i(\theta)\|^3\right) &\le N^{-3/2}\sup_{\theta\in K}[\![\sqrt{N}I_N(\theta)^{-1/2} - I(\theta)^{-1/2}]\!]\sum_{i=1}^N \mathbb{E}_\theta\left(\left\|S^i(\theta)\right\|^3\right) \\
&\quad + N^{-3/2}\sup_{\theta\in K}[\![I(\theta)^{-1/2}]\!]\sum_{i=1}^N \mathbb{E}_\theta\left(\|S^i(\theta)\|^3\right) \\
&\le CN^{-1/2}\left[\sup_{\theta\in K}[\![\sqrt{N}I_N(\theta)^{-1/2} - I(\theta)^{-1/2}]\!] + \sup_{\theta\in K}[\![I(\theta)^{-1/2}]\!]\right],
\end{aligned}$$

which converges to 0 as $N \to \infty$.

To verify (B.2), we show that (recall that $J^i(\theta)$ denotes the second derivative of $\sqrt{p^i(\theta)}$)

$$\sup_{\|h\| \leq R} \frac{1}{N} \left[\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\nu^i} \left(\left[\!\left[J^i(\theta_N) - J^i(\theta) \right]\!\right]^2 \right) \right] \quad \text{and} \quad \frac{1}{N} \left[\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\nu^i} \left(\left[\!\left[J^i(\theta) \right]\!\right]^2 \right) \right] \tag{6}$$

converge to 0 uniformly in K. As $J^i(\theta)$ is continuous, it is uniformly continuous on compacta, such that for all $i \in \mathbb{N}$, $a_{i,N} = \sup_{\|h\| \leq R} \left[\!\left[J^i(\theta_N) - J^i(\theta) \right]\!\right]$ converges a.s. to 0 as $N \to \infty$. One can show that $a_{i,N} \leq A^i(\theta, R)$ and $\mathbb{E}_{\nu^i}\left(A^i(\theta,R)^2\right) \leq D_K$. Dominated convergence implies $\mathbb{E}_\theta(a_{i,N}) \to 0$, and the uniform (in i) bound D_K implies uniform in K convergence of the left term in (6) to 0. For the right term in (6) we note that $\mathbb{E}_{\nu^i}\left(\left[\!\left[J^i(\theta)\right]\!\right]^2\right) \leq \mathbb{E}_\theta\left(\left[\!\left[\frac{d}{d\theta} S^i(\theta)\right]\!\right]^2\right) + \mathbb{E}_\theta\left(\left[\!\left[S^i(\theta)' S^i(\theta)\right]\!\right]^2\right) < C_K$, where C_K is a constant that only depends on K and we conclude uniform in $\theta \in K$ convergence to 0, completing the verification of (B.2).

2.3.2. On the convergence of the average Fisher information in SDMEMs

As seen above, a key condition for establishing the asymptotic normality of the MLEs was the convergence of the scaled N-sample Fisher information $\frac{1}{N} I_N(\theta) = \frac{1}{N} \sum_{i=1}^{N} I^i(\theta)$ to a deterministic limit $I(\theta)$. This is difficult to check when the drift contains subject-specific covariate information D^i and these covariates are not identical across subjects, because the processes X^i do not have the same distributions, since the drift function F varies across subjects, $F^i(x, \mu, \phi^i) = F(x, D^i, \mu, \phi^i)$.

In a linear regression model with random effects, the asymptotic behavior of the averaged FI is deduced from a comparable asymptotic behavior of the averaged covariates, such that the verification of conditions can conveniently be accomplished on covariate level. Also in SDMEMs with covariates, it would be desirable to be able to break down the convergence of $\frac{1}{N} I_N(\theta)$ to an average covariate behavior. This, however, is not possible, not even if we assume the simplest case where the drift function F is linear in state, covariates, fixed and random effects and if the latter are Gaussian distributed with known covariance matrix.

We illustrate this in the simplest non-trivial example that includes covariates. We look at a one-dimensional state process X^i governed by $dX_t^i = \left[X_t^i(\mu^1 + \phi^{i,1}) + D_t^i(\mu^2 + \phi^{i,2})\right] dt + dW_t^i$, with fixed effects vector $\mu = (\mu^1, \mu^2)'$, i.i.d. $\mathcal{N}(0, \Omega)$-distributed random effects $\phi^i = (\phi^{i,1}, \phi^{i,2})'$, and known covariate process D^i. We assume Ω is known, such that $\theta = \mu$ is the only unknown parameter. This setup is a special case of (2) with $A = 0$, $B = C$ and therefore $U_i := U_{1i} = U_{2i}, V_i := V_{1i} = V_{2i} = Z_i$. More specifically,

$$U_i = \begin{pmatrix} \int_0^T X_t^i dX_t^i \\ \int_0^T D_t^i dX_t^i \end{pmatrix} \quad \text{and} \quad V_i = \begin{pmatrix} \int_0^{T^i} (X_t^i)^2 dt & \int_0^{T^i} X_t^i D_t^i dt \\ \int_0^{T^i} X_t^i D_t^i dt & \int_0^{T^i} (D_t^i)^2 dt \end{pmatrix}.$$

The FI is by definition $I^i(\mu) = \mathbb{E}_\mu\left(-\frac{d^2}{d\mu^2} \log p^i(\theta)\right) = \mathbb{E}_\mu\left((I + V_i\Omega)^{-1} V_i\right)$, see eq. (4). The matrix $(I + V_i\Omega)^{-1} V_i$ is, however, a non-linear function of V_i and thus finding an explicit expression for $I^i(\mu)$ is generally impossible - even in the simple linear case, where X^i is nothing but a Gaussian process. For comparison, in the linear mixed effects model, the log-likelihood for observation y^i with covariate vectors x^i, z^i is proportional to $-\frac{1}{2}(y^i - (x^i)'\mu)'(I + z_i'\Omega z_i)^{-1}(y^i - (x^i)'\mu)$, and therefore the FI is $\mathbb{E}_\mu\left(x^i(I + z_i'\Omega z_i)^{-1}(x^i)'\right) = x^i(I + z_i'\Omega z_i)^{-1}(x^i)'$. The crucial difference, as compared to the linear SDMEM case, is that the matrix $(I + z_i'\Omega z_i)^{-1}$ is deterministic. Therefore, convergence of $\frac{1}{N}\sum_{i=1}^{N} I^i(\theta)$ is implied by a limiting behavior of averages. This is particularly attractive as one can often design the experiment in such a way that the required limiting behavior

holds. In the SDMEM case, however, it will generally not be possible to determine from an analytical expression of $I_N(\theta)$, whether the condition $\frac{1}{N} I_N(\theta) \to I(\theta)$ holds, due to the combination of non-linearity and stochasticity.

2.4. Hypothesis testing

It is commonly of interest to test whether an applied treatment has a significant effect on the treated subjects, i.e., to test whether an underlying treatment effect β, an ℓ-dimensional subparameter of the fixed effect μ, $1 \leq \ell \leq p$, is significantly different from 0. The asymptotic normality of the MLE for the SDMEMs lends itself naturally to the application of Wald tests, which can be used to investigate two-sided null hypotheses such as $H_0 : \beta = 0$ (no treatment effect) or more generally, any k-dimensional, $1 \leq k \leq \ell$, linear null hypothesis $H_0 : L\beta = \eta_0$, where L is a $k \times \ell$ matrix of rank k, specifying the linear hypotheses of interest, and $\eta_0 \in \mathbb{R}^k$. The Wald test statistic is $\left(L\hat{\beta}_N - \eta_0\right)' \left(L\hat{V}_N L'\right)^{-1} \left(L\hat{\beta}_N - \eta_0\right)$, where $\hat{\beta}_N$ is the MLE of β and $\hat{V}_N = \widehat{\mathbb{C}\mathrm{ov}}(\hat{\beta}_N)$ denotes its estimated variance-covariance matrix of $\hat{\beta}_N$. Under the null hypothesis the test statistic is asymptotically χ^2-distributed with k degrees of freedom (Lehmann and Romano, 2006). Alternatively, the Likelihood-Ratio (LR) test can be applied. Let p_0 and p_a denote the likelihoods under the null and under the alternative, then the test statistic $-2\log(p_0/p_a)$ is asymptotically χ^2-distributed with degrees of freedom equal to the difference in number of parameters. Hypothesis testing in the present SDMEM framework will be further illustrated in the following two sections.

3. Simulation study on the linear transfer model

The model under investigation is inspired from a study on the selenomethionine metabolism in humans (Große Ruse et al., 2015). This multidimensional linear transfer model finds frequent applicability in pharmacokinetics. Each component of the model's state vector represents the concentration of a substance in a certain compartment (e.g., in an organ of the human body), such that the model describes the flow between several compartments. We consider a flow in form of a cascade-shaped transfer structure, illustrated in Fig. 1. The transfer rates between compartments are mostly subject-specific, which we account for by inclusion of random effects. Additionally, dynamics may depend on covariates D^i. Here, the $D^i \in \{0, 1\}$ encodes the affinity of subject i to one of two treatment groups. For simplicity we assume a unit diffusion matrix, such that we consider the model

$$dX_t^i \;=\; F(X_t^i, D^i, \mu, \phi^i)dt + dW_t^i \;=\; -G(\alpha + \phi^i)X_t^i + D^i\beta dt + dW_t^i,$$

for $0 \leq t \leq T$ and $X_0^i = 0$, where $\mu' = (\alpha', \beta')$ is the fixed parameter and

$$G(\alpha) = \begin{pmatrix} \alpha_1 & 0 & 0 & 0 & -\alpha_5 \\ -\alpha_1 & \alpha_2 & 0 & 0 & 0 \\ 0 & -\alpha_2 & \alpha_3 + \alpha_6 & 0 & 0 \\ 0 & 0 & -\alpha_3 & \alpha_4 & 0 \\ 0 & 0 & 0 & -\alpha_4 & \alpha_5 \end{pmatrix}.$$

This is a special case of the affine model (2). The (unknown) fixed effect μ has the 6-dimensional component α, which is shared across both groups (*placebo* and *treatment*) and an additional 5-dimensional component β, which describes the effect of the covariate (treatment) on a subject's dynamics. We let $\beta' = (1, 2, 3, 1, -2)$. The random effects ϕ^i are i.i.d. $\mathcal{N}(0, \Omega)$-distributed with unknown Ω, which we assume to be diagonal with entries $\mathrm{diag}(\Omega) = \left(0.5^2, 1^2, 1^2, 0.5^2, 0.3^2, 0.3^2\right)$.

With $\alpha' = (\alpha_1, \ldots, \alpha_6) = (2, 4, 3, 2, 1, 1)$, all eigenvalues of $G(\alpha)$ have positive real parts, implying that the model has a stationary solution. The processes X^i for individuals without treatment, $D^i = 0$, are (on average) mean-reverting to 0, while those for individuals in the treatment group have average long-term mean $(G(\alpha))^{-1}\beta = (7.50, 4.25, 5.00, 8.00, 14.00)'$, see also Fig. 2. The observation horizon is fixed to $T = 15$. A trajectory of $(X_t^1, \ldots, X_t^N)_{0 \le t \le T}$ is simulated with the Euler-Maruyama scheme with simulation step size $\delta = 10^{-4}$. Fig. 2 shows four realized trajectories of the 5-dim. process X^i. The upper two panels show trajectories for $D^i = 0$ and the lower two correspond to $D^i = 1$.

3.1. Parameter estimation

To mitigate simulation errors, the simulated trajectories are thinned by a factor b (taking only every b-th observation). We explore the expected time-discretization bias of the estimators by repeating estimation for different thinning factors, $b \in \{10, 100\}$, which results in sampling intervals $\Delta t = \delta \cdot b = 0.001, 0.01$. To investigate estimation performance as a function of sample size, we used $N = 20$ and $N = 50$. Estimation for the considered $(\Delta t, N)$-combinations was repeated on $M = 500$ simulated data sets. Table 1 reports the sample estimates of relative biases and root mean squared errors (RMSE) of the fixed effects and of the variances of the random effects. The relative bias of $\hat{\alpha}_j$ is computed as $\frac{1}{M}\sum_{m=1}^{M} \frac{\hat{\alpha}_j^{(m)} - \alpha_j}{\alpha_j}$ and the RMSE as $\left(\frac{1}{M}\sum_{m=1}^{M}(\hat{\alpha}_j^{(m)} - \alpha_j)^2\right)^{1/2}, j = 1, \ldots, 6$, with an analogous definition for the other parameters. The first six rows in the table correspond to estimated biases and RMSEs of the shared fixed effects α_j, $j = 1, \ldots, 6$. The subsequent five rows show these metrics for the treatment effects β_j, $j = 1, \ldots, 5$, and the last six rows correspond to the metrics for the diagonal elements of Ω (i.e., the variances of the random effects). The estimation is very accurate already at sample sizes as small as $N = 20$, when the data is sampled at high frequency (here 1/0.001). Increasing the sample size to $N = 50$ does not add much to the accuracy of the estimation of the fixed effects. But it does, and not surprisingly, improve the estimation of the variances of the random effects, by up to 14 percentage points. For a lower sampling frequency of 1/0.01, estimates of the fixed effects α, β are on average biased by only about 1-2%, which is still very accurate. The variances of the random effects are estimated with an average bias of 5-9% for $N = 50$ and $\Delta t = 0.01$. Not displayed here are simulation results for low frequency observations with $\Delta t = 0.1$. Simulations have shown that estimation becomes fairly unreliable in this case. The bias due to the time-discretization of the continuous-time estimator is pronounced, with values of up to 25% for the fixed effects and up to almost 50% for the variances of the random effects. The RMSEs rise by more than 100%, as compared to the results obtained for a 10 times higher sampling frequency. If only low-frequency data are available, caution is therefore recommended and estimation should only be done on a data set that has been enlarged by imputing data in between the observation time points.

3.2. Hypothesis testing

A natural step is to test whether the treatment effect β, or a subparameter, is significantly different from 0. We estimate the false-positive rate of the Wald test (see subsection 2.4) for this model and investigate the test's power under different non-zero treatment effects. The estimated variance-covariance matrix $\hat{V}_N = \widehat{\mathbb{C}\mathrm{ov}}(\hat{\beta}_N)$ of $\hat{\beta}_N$ is obtained from $M = 500$ (separately) computed MLEs $\hat{\beta}_N^{(m)}, m = 1, \ldots, M$, where underlying data sets have been simulated under the true hypothesis (under H_0 for estimation of the false positive rate and under H_1 for power estimation). Table 1 shows that the estimation was accurate for high- and medium-frequency observations. Diagnostic plots (not shown here) reveal that the asymptotic distribution of the

Table 1. Linear transfer model. Shown are estimated relative bias and RMSE of $\hat{\alpha}, \hat{\beta}$ and diag$(\hat{\Omega})$. The sample sizes are $N = 20, 50$, and sampling intervals are $\Delta t = 0.001, 0.01$. For every combination $(N, \Delta t)$, the estimation was repeated on $M = 500$ generated data sets.

$(N,\ \Delta t)$		$(20, 0.001)$		$(50, 0.001)$		$(50, 0.01)$	
true value		rel. bias	RMSE	rel. bias	RMSE	rel. bias	RMSE
	2.00	0.003	0.116	0.001	0.079	-0.018	0.086
	4.00	0.001	0.232	-0.002	0.149	-0.024	0.172
α	3.00	0.003	0.253	0.001	0.163	-0.021	0.170
	2.00	-0.003	0.126	-0.001	0.083	-0.017	0.088
	1.00	0.003	0.074	0.001	0.047	-0.016	0.049
	1.00	-0.003	0.146	0.002	0.091	-0.008	0.091
	1.00	0.000	0.157	-0.002	0.099	-0.020	0.099
	2.00	-0.001	0.174	-0.002	0.114	-0.024	0.121
β	3.00	0.002	0.233	0.002	0.152	-0.010	0.152
	1.00	0.010	0.231	-0.001	0.148	0.014	0.146
	-2.00	0.006	0.203	0.002	0.124	-0.024	0.131
	0.25	-0.091	0.093	-0.037	0.062	-0.079	0.062
	1.00	-0.046	0.355	-0.035	0.208	-0.095	0.216
diag(Ω)	1.00	-0.073	0.343	-0.035	0.215	-0.085	0.219
	0.25	-0.035	0.097	-0.026	0.061	-0.065	0.060
	0.09	-0.045	0.035	-0.009	0.022	-0.047	0.021
	0.09	-0.181	0.055	-0.040	0.036	-0.065	0.035

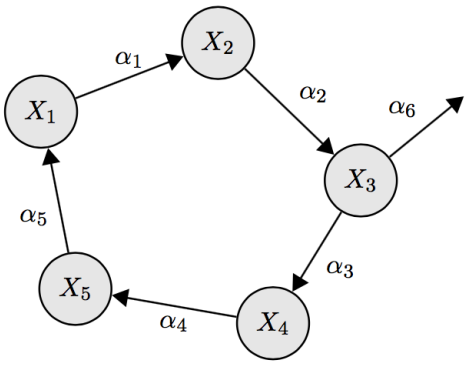


Fig. 1. Illustration of the 5-dimensional linear transfer model used in the simulation example. The state $X_j = (X_{j,t})_{0 \le t \le T}$ gives the concentration (over time) of a substance in compartment $j, j = 1, \ldots, 5$. The $\alpha_j, j = 1, \ldots, 5$, are the unknown flow rates between compartments and α_6 represents the outflow rate of the system.

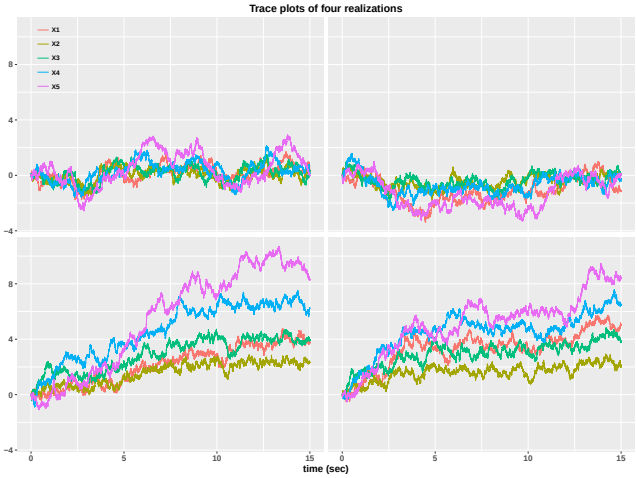


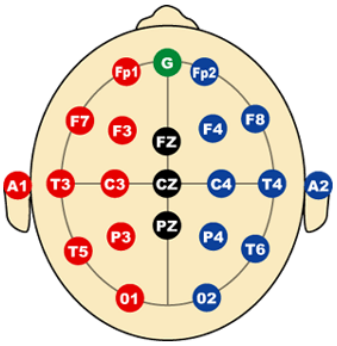
Fig. 2. Linear transfer model: Four realizations of the 5-dimensional state process. The upper two panels show realizations when the covariate is 0 (reference group) and the lower two panels display trajectories for $D^i = 1$ (treatment group). Note the clearly visible difference in the long-term means between the two groups.

MLE is close to normal already for $N = 20$ subjects, such that even for a rather small data set and a medium sampling frequency, test results are reliable. The choice $(N, \Delta t) = (20, 0.01)$ provides a simulation setting which is sufficiently reliable, but at the same time not trivial and will challenge the hypothesis test, in particular for small treatment effects. The estimated false positive rate (based on M under H_0 generated data sets) is 0.074, revealing a slightly liberal finite-sample test behavior. The power of detecting a treatment effect (rejecting $H_0 : \beta = 0$) was computed for different values of β. For $\beta = (1, 2, 3, 1, -2)'$ (values as above), the estimated power was 1. This comes to no surprise as the long-term mean $(7.5, 4.25, 5, 8, 14)'$ of the state process in the treatment group is considerably different from the zero long-term mean of the control group. The power, estimated to 0.956, was still convincing for a much smaller treatment effect $\beta = (0.1, 0.2, 0.3, 0.1, -0.2)'$, which gives a long-term mean of $(0.75, 0.425, 0.5, 0.8, 1.4)'$. This is especially impressive as the state process' standard deviation (from its long-term mean 0) under H_0 is about $(0.66, 0.49, 0.59, 0.72, 1.21)'$. More challenging is the rejection of H_0 when the treatment has a small effect on, e.g., only one coordinate, $\beta = (0.1, 0, 0, 0, 0)'$. In this case (long-term mean $(0.2, 0.1, 0.1, 0.15, 0.3)'$), and for $N = 20$ the chance of rejecting H_0 is as small as 16% and it is thus hardly possible to detect a difference between groups. However, while being only slightly conservative, the asymptotic Wald test is able to detect a treatment effect for a rather small data set, even if it causes only a little change of the long-term mean as compared to the standard deviation of the process.

4. Analysis of EEG data

Scalp electroencephalography (EEG) is a non-invasive method to measure electrical activity in the brain over time, recorded by electrodes placed on the scalp. Abnormal patterns in the recorded brain waves are used as possible indicators for diseases such as epilepsy and can help determining a suitable treatment for the patient. The data set was collected during a study conducted by the Children's Hospital Boston and is described in Shoeb (2009). It consists of continuous EEG recordings on 23 epilepsy patients. The electrodes were arranged on the scalp according to the international 10-20 system (see Fig. 3) and the EEG signal was recorded with a sampling frequency of 256 Hz. This is high frequency compared to the typical time scales of the system, and thus, for this type of data, the discretization error will be negligible. During time of recording, every patient experienced one or more periods of abnormal activity that have been classified as epileptic seizures by Shoeb (2009).

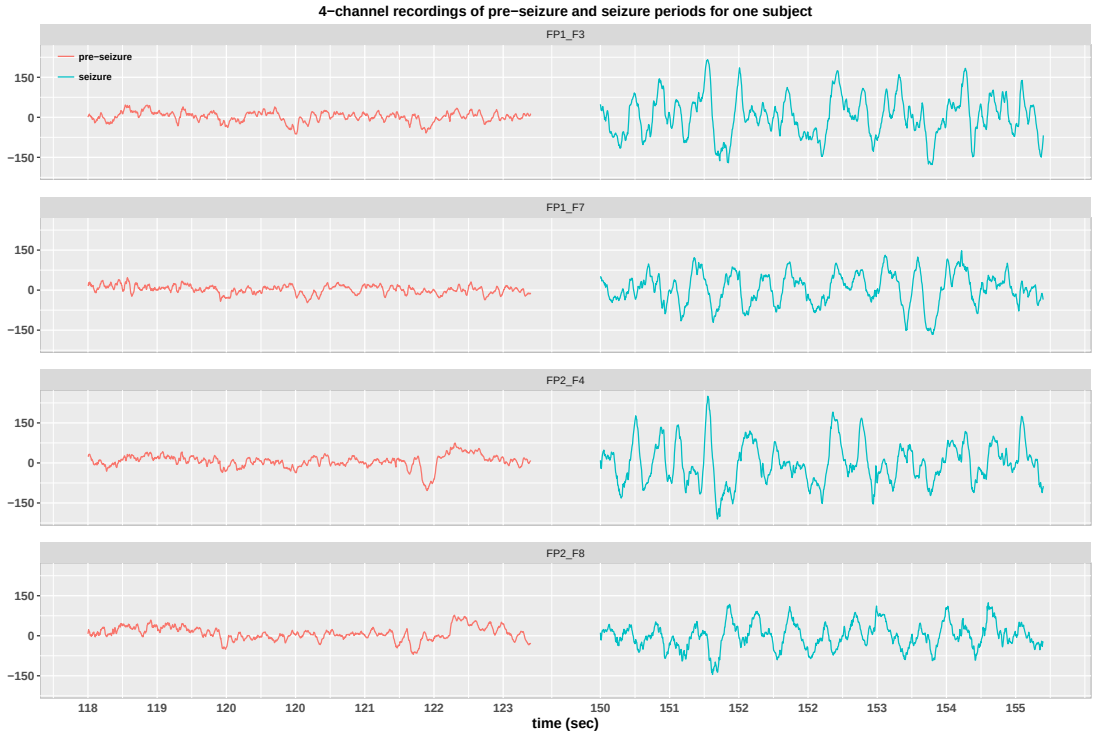
Fig. 3. Location of scalp electrodes according to the international 10-20 system



Part of this data set was also analyzed in Østergaard et al. (2017). Their results, which were obtained using a different modeling approach, indicated increased channel interaction strength during seizure. However, their findings were based on data from a single subject only. It is therefore of interest whether one can infer an increased interaction when combining data from several subjects within a dynamical mixed-effects framework. We focus our analysis on recordings from four channels in the frontal lobe, FP1_F7, FP1_F3, FP2_F4 and FP2_F8, as done in Østergaard et al. (2017). Thus, responses are four-dimensional time series for every patient. The first two channels are located on the left hemisphere and the latter two are, mirrored, on the right. For every patient we extracted two $5s$ periods of recording, one of them reflecting normal brain activity and the other reflecting abnormal activity classified as epileptic seizure. Fig. 4 shows data for the selected periods of an 11-year old boy. The rows in the plot correspond to the four selected channels and the color indicates pre-seizure (red) and seizure (blue) sections.

The dynamics of the signals during seizure differ clearly from pre-seizure behavior and the objective of this analysis is to better understand, quantitatively and qualitatively, how they differ. From a neurophysiological viewpoint the interaction structure between brain regions or different channels is of interest and, in particular, if and how this network structure changes under different conditions, e.g., when patients enter an epileptic seizure state. A hint on possible interactions can be obtained by investigating the correlation structure between channels. Under a sufficiently short time window, the otherwise non-stationary behavior of spontaneous brain activity can be considered stationary. We model the $5s$ sections of EEG recordings from the four selected channels by a stationary four-dimensional Ornstein-Uhlenbeck (OU) process. This is a process with dynamics $dX_t = AX_t dt + \Sigma dW_t$ and explicit solution $X_t = \mathrm{e}^{At}x_0 + \int_0^t \mathrm{e}^{A(t-s)}\Sigma dW_s$. In particular, X_t (given x_0) is Gaussian with mean $\mathbb{E}(X_t) = \mathrm{e}^{At}x_0$ and covariance matrix $\mathbb{V}(X_t) = \int_0^t \mathrm{e}^{As}\Sigma\Sigma'\mathrm{e}^{A's}ds$. If all eigenvalues of the rate matrix A have negative real parts, X has a stationary solution and the stationary distribution is a centered Gaussian with covariance matrix $V = \int_0^\infty \mathrm{e}^{Au}\Sigma\Sigma'\mathrm{e}^{A'u}du$ and autocorrelation function (ACF) $r_X(\tau) = V^{1/2}\mathrm{e}^{A'\tau}V^{-1/2}$.

Fig. 4. Sections of $5s$ of the EEG-recording during pre-seizure activity and during an epileptic seizure for the four channels FP1_F7, FP1_F3, FP2_F4 and FP2_F8 and a single subject, an 11-year old boy. The pre-seizure time series is plotted in red and the signal during a subsequent seizure is given in blue.



The statistical model

The prevalent inter-subject variability for EEG data is one of the greater challenges for any inference procedure (Shoeb, 2009), and we account for such subject-specific deviations from mean OU dynamics by the inclusion of random effects. We present the subject-specific SDMEM model for the EEG data first and afterwards give a motivation for our choice. We denote the pre-seizure process of subject i by $Y^{i,1}$ and the seizure process by $Y^{i,2}$. During seizure, the signal is amplified considerably (Fig. 4). As structural differences are easier to analyze when pre-seizure and seizure data are of comparable magnitude, we re-scale the data to $X_t^{i,k} = \mathrm{diag}(1/\sigma_{11}^{i,k}, \dots, 1/\sigma_{44}^{i,k}) Y_t^{i,k}$, with $\sigma_{jj}^{i,k}$ being the infinitesimal standard deviation (square root of the quadratic variation) of channel j. Normalizing by a diagonal matrix does not introduce changes to the inherent channel structure, but only affects the scaling. The specific choice of the scaling renders the quadratic variation of the obtained processes $X^{i,k}$ to be a correlation-matrix type. Taking the OU-dynamics as base model, we then model the (normalized) data for subject i by

$$dX_t^{i,k} = \left[A + \Phi^{i,\mathrm{pre}} + D^{i,k}(M + \Phi^{i,\delta})\right] X_t^{i,k} dt + \Sigma dW_t^{i,k}, \tag{7}$$

where $W^{i,k}$ are independent Brownian motions, $A, M, \Phi^{i,\mathrm{pre}}, \Phi^{i,\delta}$ are 4×4 matrices and the entries of $\Phi^{i,\mathrm{pre}}, \Phi^{i,\delta}$ are independent centered Gaussian random variables (the random effects). The covariate $D^{i,k}$ encodes whether the data belongs to pre-seizure ($D^{i,1} = 0$) or seizure state ($D^{i,2} = 1$). Thus, for a pre-seizure state, population dynamics are driven by the rate matrix A,

whereas M represents the covariate (or seizure) effect. Rewriting equation (7) as

$$dX_t^{i,k} = \left[B(X_t^{i,k}, D^{i,k})\mu + C(X_t^{i,k}, D^{i,k}) \begin{pmatrix} \phi^{i,\mathrm{pre}} \\ \phi^{i,\delta} \end{pmatrix} \right] dt + \Sigma dW_t^{i,k}, \tag{8}$$

(with $\phi^{i,\mathrm{pre}}$ and $\phi^{i,\delta}$ being the vectorized versions of $\Phi^{i,\mathrm{pre}}$ and $\Phi^{i,\delta}$, respectively) reveals that this model belongs to the class of affine SDMEMs with covariates, model (2), and thus has explicit likelihood and fixed-effects estimators.

Motivation for the model approach

The processes $W^{i,1}$ and $W^{i,2}$ represent the noise within the system on a short time scale. Their independence is supported by the fact that data sections $X^{i,1}, X^{i,2}$ are temporally (on a larger time scale) clearly separated. In general, behavior during seizures is more variable, and, in particular, show a stronger amplification.

Fig. 5. Diagonal plots: Infinitesimal standard deviation for every channel, estimated by the square root of the quadratic variation, used to normalize the data records. Off-diagonal plots: Infinitesimal correlation between channels. The gray lines correspond to individual estimates for pre-seizure (red) and seizure (blue) data. The black line is the mean correlation, averaged over individuals and states.

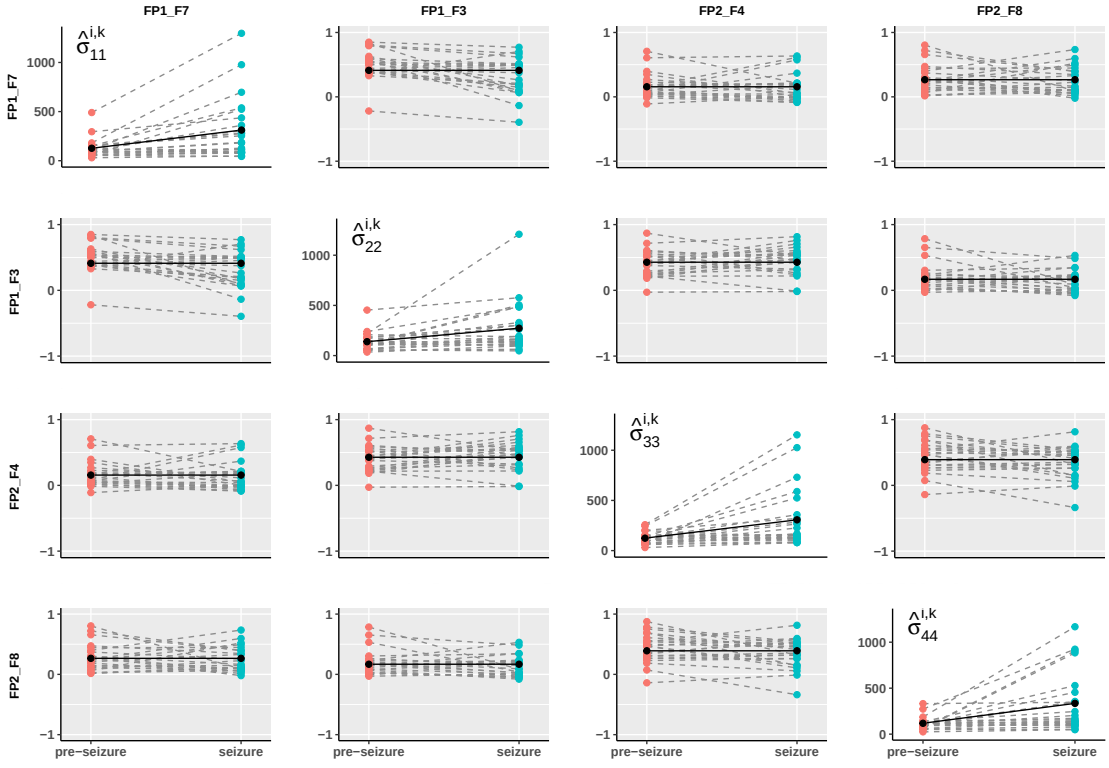


Fig. 5 shows that the average structure of the infinitesimal correlations between channels (off-diagonal plots) does not differ considerably between pre-seizure (red) and seizure states

(blue). The estimated infinitesimal standard deviations $\hat{\sigma}_{jj}^{i,k}$ of the channels (diagonal plots) reveal, however, that in most subjects and channels (80%) the standard deviation increases, in the most extreme case it increases 14-fold, and in 78% of the cases it more than doubles. Because of the shared infinitesimal correlation structure we model the normalized pre-seizure and seizure processes with the same diffusion matrix, denoted above by Σ. This implies that any further changes apart from the scaling in the dynamics between states are captured by changes in the drift. The transition from pre-seizure to seizure state is modeled in terms of the drift matrix $M+\Phi^{i,\delta}$. The structural change in the population dynamics is represented by M, and the change in the subject-specific variation due to seizure is represented by the random effect $\Phi^{i,\delta}$.

Results

The statistical conclusions are based on the population rate matrices A, M, which are estimated by their MLEs as outlined in section 2.2. The estimates of the population-based rate matrices are

$$\hat{A} = \begin{pmatrix} -10.52 & -3.59 & -0.42 & 2.47 \\ 3.24 & -17.72 & 4.76 & 1.70 \\ 1.98 & 0.14 & -12.60 & 3.94 \\ 0.74 & -1.75 & -1.52 & -12.87 \end{pmatrix}; \quad \hat{M} = \begin{pmatrix} -3.22 & 2.65 & 0.80 & -0.16 \\ 0.83 & 4.60 & -1.51 & 2.81 \\ -0.82 & -0.27 & 0.74 & 0.00 \\ 3.27 & 0.59 & 1.30 & -3.36 \end{pmatrix}.$$

The eigenvalues of $\hat{A}$ and $\hat{A}+\hat{M}$ have negative real parts, such that stationary distributions on the population level for pre-seizure and seizure states indeed exist.

Fig. 6. Confidence intervals for the estimated entries of the covariate effect matrix M. The blue lines are for the full model, whereas the 16 black lines are derived from 16 reduced models where all but one entry of M are set to 0.

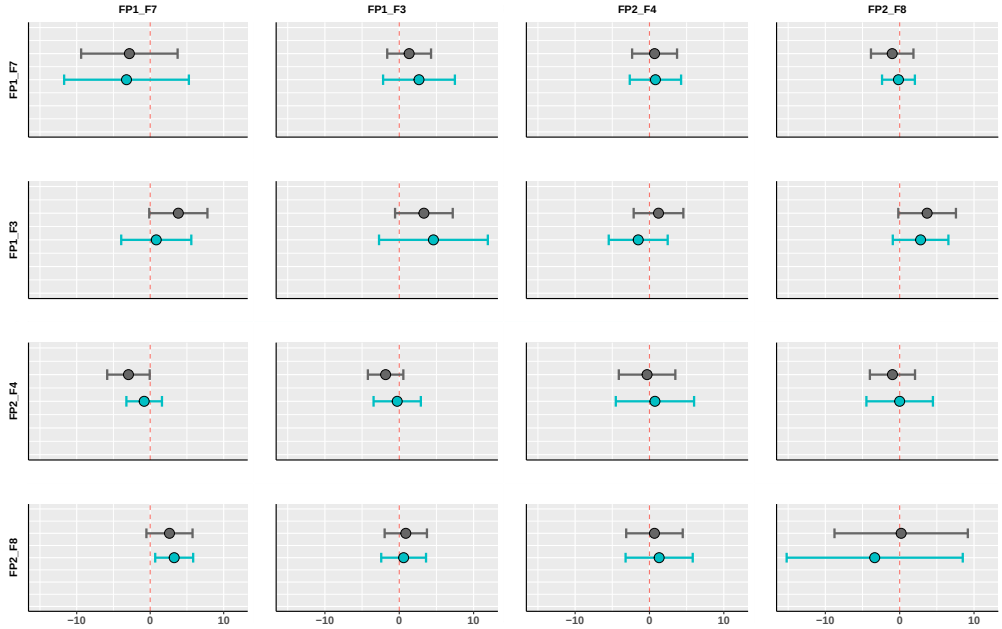


Table 2. Stationary correlations between channels, for pre-seizure state and seizure state (columns 4,5). The last column shows the change in correlation for seizure epochs as compared to non-seizure periods.

channel 1	**channel 2**	**correlation**		**change** (%)
		pre-seizure	*seizure*	
FP1_F7	FP1_F3	0.42	0.52	23.81
FP1_F7	FP2_F4	0.22	0.26	18.18
FP1_F7	FP2_F8	0.32	0.43	34.37
FP1_F3	FP2_F4	0.60	0.59	-1.67
FP1_F3	FP2_F8	0.23	0.36	56.52
FP2_F4	FP2_F8	0.43	0.47	9.30

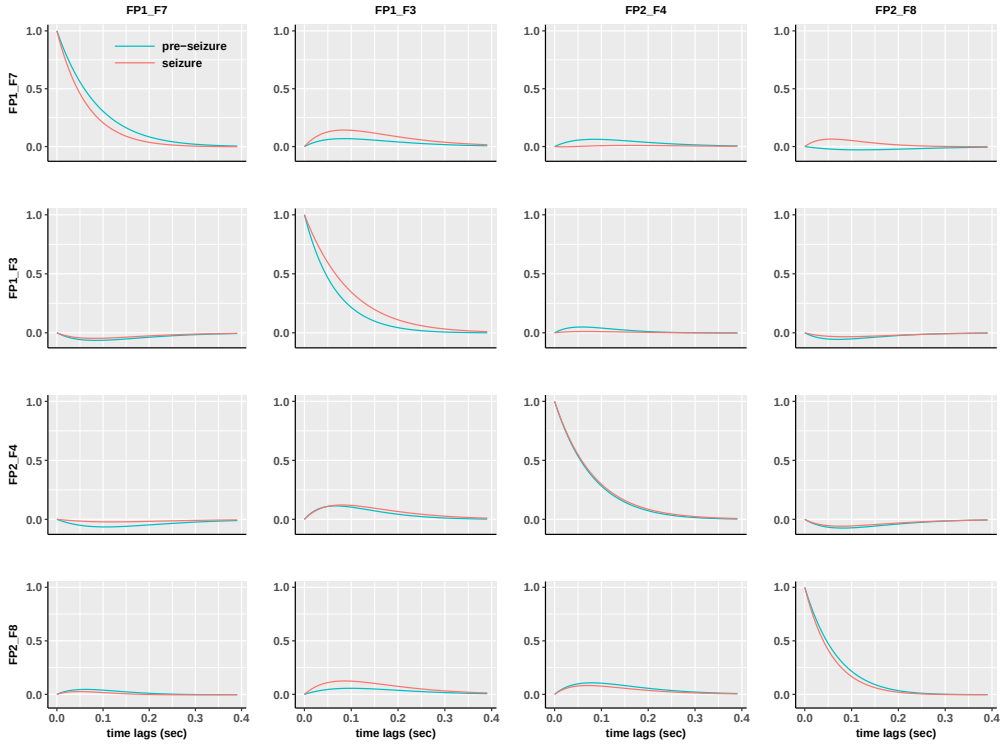
In a first step we assess whether the overall covariate effect M is significant by testing $H_0 : M = 0$ versus $H_0 : M \neq 0$ with a likelihood ratio test. The likelihood ratio statistic, which under H_0 is asymptotically χ^2_{32-16}-distributed, has a realized value of 13.71, with a p-value of 0.62. We conclude that the null hypothesis $H_0 : M = 0$ cannot be rejected on a 5%-level. However, the data set consists of observations from only 23 subjects and the number of fixed effects alone (32 parameters) is considerably higher. Therefore, a possible prevalent covariate effect is hard to detect and statistical results have to be interpreted with caution. More insight into where changes might be present in the rate matrix between pre-seizure and seizure states is provided in Fig. 6. It shows the 95%-confidence intervals (CIs) for every entry of M in blue. Only one element of M has a CI that does not include 0. A way to increase statistical power is to cut down on the number of unknown parameters. Considering only one element of M active instead of all 16, the number of unknown fixed effects is reduced from 32 to 17. Each of the black CIs in Fig. 6 is derived from a reduced model in which all but one elements of M are set to 0. As expected, most CIs are more narrow, however, only few elements appear to have an effect. The lower left plot, e.g., suggests an increased influence of channel FP2_F8 on FP1_F7 under seizure.

It is not straightforward to interpret a covariate effect by looking at the matrix M in an entry-by-entry manner. Insights about structural changes in the underlying dynamics can more easily be gained by looking at interactions in the system. Interactions can be assessed by the correlations between components of $X_t^{i,k}$. To analyze this, we compare the (population) stationary covariance matrices of pre-seizure and seizure state, which will reveal differences in the long-run correlation structure between channels. The population estimates of the correlation matrices of the stationary distributions for pre-seizure and seizure states are shown in Table 2. In line with the findings in Østergaard et al. (2017), channel-correlations increase during seizure, most of them by at least around 20%.

Other quantities of interest are the ACFs shown in Fig. 7. The diagonal panels in Fig. 7 show the univariate autocorrelation for every channel, i.e., the correlations between a channel and its time-lagged version, as a function of the time lag. The autocorrelations show no marked difference between pre-seizure and seizure states. This can also be summarized by the eigenvalues of matrices $\hat{A}$ and $\hat{A} + \hat{M}$. The absolute values of the real parts provide the rates of decay, and thus, their inverse indicate the typical time constants in the system. For the pre-seizure state the absolute values of the real parts vary between 11.4 and 17.7, whereas during seizure these vary between 11.6 and 18.0.

To summarize, despite not being statistically significant, there are indications of changes in the correlation structures during epileptic seizures, where correlations between channels increase.

Fig. 7. Theoretical multivariate ACFs $r_X^k(\tau) = \hat{V}_k^{1/2} e^{\hat{A}_k' \tau} \hat{V}_k^{-1/2}, k = 1, 2$, for the stationary distributions, where $\hat{A}_1 = \hat{A}$ is the estimated population rate matrix for pre-seizure states (in blue), $\hat{A}_2 = \hat{A} + \hat{M}$ is for seizure states (red), and $\hat{V}_k$ is the estimated stationary (population) covariance matrix.



Nonetheless, the main effect of an epileptic seizure seems to be captured by increased variance addressed in the rescaling of the data, more than structural changes. However, with only 23 patients finer effects might be difficult to unravel. An analysis of all channels would be of interest, but is only possible with much larger sample sizes due to the many parameters a full analysis would imply.

5. Discussion

SDMEMs constitute an attractive class of statistical models for biomedical data. We suggested an approach for parameter inference in this framework, which even comprises more complex dynamics such as time-inhomogeneity and multivariate and nonlinear states. The inclusion of (deterministic) subject-specific covariate information, which causes the modeler to leave the world of identically distributed observations, is addressed as well. The presented conditions for consistency and asymptotic normality of the MLE along the lines of L_2-differentiability do not require the typical strong smoothness properties of densities and thereby open doors to irregular models. To make abstract formulations graspable, conditions are illustrated for the special case of Gaussian random effects and linear parameters (but possible non-linearity in the state). This model is a multidimensional extension of the one studied in Delattre et al. (2013) and is, in its multidimensional version, particularly interesting as it comprises numerous well-

known models. Among them are the predator-prey (or Lotka-Volterra) model (Murray, 2002), the Lorenz equations introduced by Lorenz (1963), which have been used to model, e.g., temperature, wind speed and humidity, the Brusselator model (Kondepudi and Prigogine, 2014, 19.4), the Fitzhugh-Nagumo model (FitzHugh, 1955; Nagumo et al., 1962; Jensen et al., 2012), which is used to describe the regenerative firing mechanism in an excitable neuron, and the SIR (susceptible-infected-removed) model introduced by Kermack and McKendrick (1927), an epidemic model which has been widely studied and applied (Keeling and Rohani, 2008; Guy et al., 2015).

The estimation quality in terms of sample size and sampling frequency was investigated in a simulation study for a popular model in pharmacokinetics, which was motivated by a recent study (Große Ruse et al., 2015). It includes subject-specific covariate information and is linear in parameters and state. When observations are sampled at high frequency, estimation results were convincing already for small sample sizes ($N = 20$), despite the comparably large number (11 fixed effects and 6 variances) of unknown parameters. A moderate sampling interval (of $\Delta t = 0.01$) still gave good results for the considered sample size. However, when sampling at low frequency ($\Delta t = 0.1$), the discrete-time bias makes itself felt (not included here). The asymptotic normality of the MLE lends itself naturally to hypothesis tests by means of the Wald or the LR test. Based on the simulated data, we estimate the false-positive rate, revealing a slight liberalism of the test procedure, and compute the test's power for different true values of parameters.

Finally, we apply the framework to the statistical analysis of epileptic EEG data to assess differences between dynamics for non-seizure and seizure periods. The population voltage dynamics during non-seizure and seizure states are modeled as Ornstein-Uhlenbeck processes, while the prevalent inter-subject variability was accounted for by the inclusion of random effects in the drift. After having adjusted for the subject-specific deviations, systematic differences between pre-seizure and seizure recordings are assessed by comparing the population correlation structure of the corresponding stationary distributions. Our findings support those in Østergaard et al. (2017), which indicate increased state (channel) correlation for seizure epochs as compared to non-seizure states.

A few comments are in order concerning the simulation study and the presented application. Regarding the EEG data analysis, it should be noted that a physiological interpretation of our results in terms of an underlying network structure has to be taken with a grain of salt for two key reasons. One is that EEG recordings are only proxies for underlying brain activity. Secondly, correlation is only one way to assess signal interaction. Non-linear interactions, which are undetectable by correlation-based measures, may still exist. In terms of our simulation settings, we have only studied the method's applicability to models with up to 17 parameters. Even in the case of an explicit likelihood, the MLE of the (unknown) covariance matrix of the random effects vector is implicit and estimation requires numerical optimization, which may hamper estimation when the parameter space has a high dimension.

A drawback of the presented approach is the already mentioned inherent discrete-time bias of the estimation procedure. It is negligible if observations are sampled at sufficiently high frequency, such as for EEG recordings, but for low-frequency observations, which sometimes occur in pharmacological applications, a severe bias is introduced, which is to bear in mind in applications. A possible solution is to impute data at time points in between observation times, and conduct the estimation on the enlarged data set (Bladt et al., 2016). Related to that is the problem of incomplete observations, where only some of the coordinates in the state space are observed, and an entire path of a completely unobserved (latent) coordinate should be inferred (Berg and Ditlevsen, 2013; Ditlevsen and Samson, 2014). Missing observations of one or more coordinates is not untypical for biological data. This, at a first step, prohibits application of the proposed estimation procedure, as it relies on the assumption of complete data observations. Such statistical recovery of hidden state coordinates remains a topic for future research.

Acknowledgements

The work is part of the Dynamical Systems Interdisciplinary Network, University of Copenhagen. Adeline Samson has been partially supported by the LabExPERSYVAL-Lab (ANR-11-LABX-0025-01).

References

Aït-Sahalia, Y. (2002). Maximum likelihood estimation of discretely sampled diffusions: A closed-form approximation approach. *Econometrica 70*(1), 223–262.

Beal, S. L. and L. B. Sheiner (1981). Estimating population kinetics. *Critical Reviews in Biomedical Engineering 8*(3), 195–222.

Berg, R. W. and S. Ditlevsen (2013). Synaptic inhibition and excitation estimated via the time constant of membrane potential fluctuations. *Journal of Neurophysiology 110*(4), 1021–1034.

Bladt, M., S. Finch, and M. Sørensen (2016). Simulation of multivariate diffusion bridges. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78*(2), 343–369.

Bradley, R. A. and J. J. Gart (1962). The asymptotic properties of ML estimators when sampling from associated populations. *Biometrika 49*(1/2), 205–214.

Delattre, M., V. Genon-Catalot, and C. Laredo (2017). Parametric inference for discrete observations of diffusion processes with mixed effects. *Stochastic Processes and their Applications*.

Delattre, M., V. Genon-Catalot, and A. Samson (2013). Maximum likelihood estimation for stochastic differential equations with random effects. *Scandinavian Journal of Statistics 40*(2), 322–343.

Delattre, M., V. Genon-Catalot, and A. Samson (2015). Estimation of population parameters in stochastic differential equations with random effects in the diffusion coefficient. *ESAIM: Probability and Statistics 19*, 671–688.

Delyon, B., V. Lavielle, and E. Moulines (1999). Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics 27*(1), 94–128.

Ditlevsen, S. and A. Samson (2014). Estimation in the partially observed stochastic Morris–Lecar neuronal model with particle filter and stochastic approximation methods. *The Annals of Applied Statistics 8*(2), 674–702.

Donnet, S., J.-L. Foulley, and A. Samson (2010). Bayesian analysis of growth curves using mixed models defined by stochastic differential equations. *Biometrics 66*(3), 733–741.

Durham, G. B. and A. R. Gallant (2002). Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. *Journal of Business & Economic Statistics 20*(3), 297–338.

Fahrmeir, L. and H. Kaufmann (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, 342–368.

FitzHugh, R. (1955). Mathematical models of threshold phenomena in the nerve membrane. *The Bulletin of Mathematical Biophysics 17*(4), 257–278.

Große Ruse, M., L. R. Søndergaard, S. Ditlevsen, M. Damgaard, S. Fuglsang, J. T. Ottesen, and J. L. Madsen (2015). Absorption and initial metabolism of 75 se-l-selenomethionine: a kinetic model based on dynamic scintigraphic data. *British Journal of Nutrition 114*(10), 1718–1723.

Guy, R., C. Larédo, and E. Vergu (2015). Approximation of epidemic models by diffusion processes and their statistical inference. *Journal of Mathematical Biology 70*(3), 621–646.

Hoadley, B. (1971). Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case. *The Annals of Mathematical Statistics 42*, 1977–1991.

Ibragimov, I. A. and R. Z. Has'minskii (2013). *Statistical Estimation: Asymptotic Theory*, Volume 16. Springer Science.

Jensen, A. C., S. Ditlevsen, M. Kessler, and O. Papaspiliopoulos (2012). Markov chain Monte Carlo approach to parameter estimation in the Fitzhugh-Nagumo model. *Physical Review E 86*(4), 041114.

Keeling, M. J. and P. Rohani (2008). *Modeling infectious diseases in humans and animals*. Princeton University Press.

Kermack, W. O. and A. G. McKendrick (1927). A contribution to the mathematical theory of epidemics. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, Volume 115, pp. 700–721. The Royal Society.

Kessler, M., A. Lindner, and M. Sørensen (2012). *Statistical methods for stochastic differential equations*. CRC Press.

Kondepudi, D. and I. Prigogine (2014). *Modern thermodynamics: from heat engines to dissipative structures*. John Wiley &; Sons.

Le Cam, L. (2012). *Asymptotic methods in statistical decision theory*. Springer Science, New York.

Leander, J., T. Lundh, and M. Jirstrand (2014). Stochastic differential equations as a tool to regularize the parameter estimation problem for continuous time dynamical systems given discrete time measurements. *Mathematical Biosciences 251*, 54–62.

Lehmann, E. L. and J. P. Romano (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

Lo, A. W. (1988). Maximum likelihood estimation of generalized Itô processes with discretely sampled data. *Econometric Theory 4*(2), 231–247.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences 20*(2), 130–141.

Møller, J. B., R. V. Overgaard, H. Madsen, T. Hansen, O. Pedersen, and S. H. Ingwersen (2010). Predictive performance for population models using stochastic differential equations applied on data from an oral glucose tolerance test. *Journal of Pharmacokinetics and Pharmacodynamics 37*(1), 85–98.

Murray, J. D. (2002). *Mathematical Biology I: An Introduction*, Volume 17 of *Interdisciplinary Applied Mathematics*. Springer, New York, NY, USA,.

Nagumo, J., S. Arimoto, and S. Yoshizawa (1962). An active pulse transmission line simulating nerve axon. *Proceedings of the IRE 50*(10), 2061–2070.

Østergaard, J., A. Rahbek, and S. Ditlevsen (2017). Oscillating systems with cointegrated phase processes. *Journal of Mathematical Biology 75*(4), 845–883.

Pedersen, A. R. (1995). A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scandinavian Journal of Statistics 22*, 55–71.

Phillips, P. C. and J. Yu (2009). Maximum likelihood and gaussian estimation of continuous time models in finance. In *Handbook of Financial Time Series*, pp. 497–530. Springer, New York.

Picchini, U., A. De Gaetano, and S. Ditlevsen (2010). Stochastic differential mixed-effects models. *Scandinavian Journal of Statistics 37*(1), 67–90.

Picchini, U. and S. Ditlevsen (2011). Practical estimation of high dimensional stochastic differential mixed-effects models. *Computational Statistics & Data Analysis 55*(3), 1426–1444.

Picchini, U., S. Ditlevsen, A. De Gaetano, and P. Lansky (2008). Parameters of the Diffusion Leaky Integrate-and-Fire Neuronal Model for a Slowly Fluctuating Signal. *Neural Computation 20*(11), 2696–2714.

Shoeb, A. H. (2009). *Application of machine learning to epileptic seizure onset detection and treatment*. Ph. D. thesis, Massachusetts Institute of Technology.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge University Press.

Wolfinger, R. (1993). Laplace's approximation for nonlinear mixed models. *Biometrika 80*(4), 791–795.

Bibliography

[1] Monolix version 2016R1, 2016. URL `http://lixoft.com/products/monolix/`.

[2] Y. Aït-Sahalia. Maximum likelihood estimation of discretely sampled diffusions: A closed-form approximation approach. *Econometrica*, 70(1):223–262, 2002.

[3] H. Amann and J. Escher. *Analysis III*. Birkhäuser Basel, 2009.

[4] M. Bachar, J. J. Batzel, and S. Ditlevsen, editors. *Stochastic Biomathematical Models: with Applications to Neuronal Modeling*, volume 2058 of *Lecture Notes in Mathematics, Mathematical Biosciences Subseries*. Springer, Heidelberg, New York, 2012.

[5] R. C. Berwick, K. Okanoya, G. J. Beckers, and J. J. Bolhuis. Songs to syntax: the linguistics of birdsong. *Trends in Cognitive Sciences*, 15(3):113–121, 2011.

[6] A. Beskos, O. Papaspiliopoulos, G. O. Roberts, and P. Fearnhead. Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *Journal of the Royal Statistical Society: Series B*, 68(3):333–382, 2006.

[7] B. Boashash. *Time-Frequency Signal Analysis and Processing: A Comprehensive Reference*. Elsevier, London, 2003.

[8] M. S. Brainard and A. J. Doupe. What songbirds teach us about learning. *Nature*, 417(6886): 351–358, 2002.

[9] M. S. Brainard and A. J. Doupe. Translating birdsong: Songbirds as a model for basic and applied medical research. *Annual Review of Neuroscience*, 36:489–517, 2013.

[10] C. K. Catchpole. Song repertoires and reproductive success in the great reed warbler Acrocephalus arundinaceus. *Behavioral Ecology and Sociobiology*, 19(6):439–445, 1986.

[11] C. K. Catchpole and A. Rowell. Song sharing and local dialects in a population of the european wren Troglodytes troglodytes. *Behaviour*, 125(1/2):67–78, 1993.

[12] C. K. Catchpole and P. J. Slater. *Bird Song: Biological Themes and Variations.* Cambridge Univ. Press, Cambridge, UK, 2008.

[13] C. Darwin. The descent of man. *Reprinted in Penguin Classics Series,* 2004, 1871.

[14] M. Delattre and M. Lavielle. Coupling the SAEM algorithm and the extended Kalman filter for maximum likelihood estimation in mixed-effects diffusion models. *Statistics and its Interface*, 6 (4):519–532, 2013.

[15] M. Delattre, V. Genon-Catalot, and A. Samson. Maximum likelihood estimation for stochastic differential equations with random effects. *Scandinavian Journal of Statistics*, 40(2):322–343, 2013.

[16] M. Delattre, V. Genon-Catalot, and A. Samson. Estimation of population parameters in stochastic differential equations with random effects in the diffusion coefficient. *ESAIM: Probability and Statistics*, 19:671–688, 2015.

[17] B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics*, 27(1):94–128, 1999.

[18] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, pages 1–38, 1977.

[19] S. Donnet and A. Samson. Parametric inference for mixed models defined by stochastic differential equations. *ESAIM Probability and Statistics*, 12:196–218, 2008.

[20] S. Donnet, J.-L. Foulley, and A. Samson. Bayesian analysis of growth curves using mixed models defined by stochastic differential equations. *Biometrics*, 66(3):733–741, 2010.

[21] G. B. Durham and A. R. Gallant. Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. *Journal of Business & Economic Statistics*, 20(3):297–338, 2002.

[22] J. J. Dziak, R. Li, X. Tan, S. Shiffman, and M. P. Shiyko. Modeling intensive longitudinal data with mixtures of nonparametric trajectories and time-varying effects. *Psychological Methods*, 20 (4):444, 2015.

[23] C. P. Elemans, J. H. Rasmussen, C. T. Herbst, D. N. Düring, S. A. Zollinger, H. Brumm, K. H. Srivastava, N. Svane, M. Ding, O. N. Larsen, S. J. Sober, and J. G. Švec. Universal mechanisms of sound production and control in birds and mammals. *Nature Communications*, 6:8978, 2015.

[24] O. Elerian, S. Chib, and N. Shephard. Likelihood inference for discretely observed nonlinear diffusions. *Econometrica*, 69(4):959–993, 2001.

[25] B. Eraker. MCMC analysis of diffusion models with application to finance. *Journal of Business & Economic Statistics*, 19(2):177–191, 2001.

[26] Y. O. Espmark, H. M. Lampe, and T. K. Bjerke. Song conformity and continuity in song dialects of redwings Turdus iliacus and some ecological correlates. *Ornis Scandinavica*, 20(1):1–12, 1989.

[27] R. FitzHugh. Mathematical models of threshold phenomena in the nerve membrane. *The Bulletin of Mathematical Biophysics*, 17(4):257–278, 1955.

[28] D. Florens-Zmirou. Approximate discrete-time schemes for statistics of diffusion processes. *Statistics: A Journal of Theoretical and Applied Statistics*, 20(4):547–557, 1989.

[29] I. I. Gikhman and A. V. Skorokhod. *The theory of stochastic processes 3*. Springer, 1979.

[30] A. Golightly and D. J. Wilkinson. Bayesian inference for nonlinear multivariate diffusion models observed with error. *Computational Statistics & Data Analysis*, 52(3):1674–1693, 2008.

[31] D. J. Goodfellow, P. J. Slater, and F. A. Clements. Local and regional variations in chaffinch song and the question of dialects. *Behaviour*, 88(1-2):76, 1984.

[32] M. Hansson-Sandsten, M. Tarka, J. Caissy-Martineau, B. Hansson, and D. Hasselquist. SVD-based classification of bird singing in different time-frequency domains using multitapers. In *19th European Signal Processing Conference*, pages 966–970. IEEE, 2011.

[33] D. Hasselquist. Polygyny in great reed warblers: A long-term study of factors contributing to male fitness. *Ecology*, 79(7):2376–2390, 1998.

[34] D. Hasselquist, S. Bensch, and T. von Schantz. Correlation between male song repertoire, extra-pair paternity and offspring survival in the great reed warbler. *Nature*, 381:229–232, 1996.

[35] A. L. Hodgkin and A. F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 117(4):500, 1952.

[36] A. Jarne, D. Commenges, L. Villain, M. Prague, Y. Lévy, and R. Thiébaut. Modeling $CD4^+$ T cells dynamics in HIV-infected patients receiving repeated cycles of exogenous interleukin 7. *Annals of Applied Statistics*, 11(3):1593–1616, 09 2017.

[37] A. C. Jensen, S. Ditlevsen, M. Kessler, and O. Papaspiliopoulos. Markov chain Monte Carlo approach to parameter estimation in the Fitzhugh-Nagumo model. *Physical Review E*, 86(4): 041114, 2012.

[38] I. Karatzas and S. Shreve. *Brownian Motion and Stochastic Calculus*. Graduate Texts in Mathematics. Springer New York, 1991.

[39] S. Klim, S. B. Mortensen, N. R. Kristensen, R. V. Overgaard, and H. Madsen. Population stochastic modelling (PSM)—an R package for mixed-effects models based on stochastic differential equations. *Computer Methods and Programs in Biomedicine*, 94(3):279–289, 2009.

[40] N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38: 963–974, 1982.

[41] M. Lavielle. *Mixed effects models for the population approach: models, tasks, methods and tools*. CRC Press, 2014.

[42] J. Leander, T. Lundh, and M. Jirstrand. Stochastic differential equations as a tool to regularize the parameter estimation problem for continuous time dynamical systems given discrete time measurements. *Mathematical Biosciences*, 251:54–62, 2014.

[43] J. Leander, J. Almquist, C. Ahlström, J. Gabrielsson, and M. Jirstrand. Mixed effects modeling using stochastic differential equations: illustrated by pharmacokinetic data of nicotinic acid in obese zucker rats. *The AAPS Journal*, 17(3):586–596, 2015.

[44] G. Lindgren, H. Rootzén, and M. Sandsten. *Stationary stochastic processes for scientists and engineers*. CRC press, 2013.

[45] R. Lipster and A. N. Shiryaev. *Statistics of Random Processes: I. General Theory*. Applications of Mathematics: Stochastic Modelling and Applied Probability. Springer, 2001.

[46] X. Mao. *Stochastic differential equations and applications*. Elsevier, 2007.

[47] P. K. McGregor. Song dialects in the corn bunting (Emberiza calandra). *Zeitschrift für Tierpsychologie*, 54(3):285 – 297, 1980.

[48] E. H. Miller. Acoustic differentiation and speciation in shorebirds. In D. E. Kroodsma and E. H. Miller, editors, *Ecology and evolution of acoustic communication in birds*, pages 241–257. Cornell University Press, Ithaca and London, 1997.

[49] E. H. Miller and D. E. Kroodsma. *Ecology and evolution of acoustic communication in birds.* Comstock, Ithaca, 1996. ISBN 0801430496.

[50] J. B. Møller, R. V. Overgaard, H. Madsen, T. Hansen, O. Pedersen, and S. H. Ingwersen. Predictive performance for population models using stochastic differential equations applied on data from an oral glucose tolerance test. *Journal of Pharmacokinetics and Pharmacodynamics*, 37(1):85–98, 2010.

[51] S. B. Mortensen, S. Klim, B. Dammann, N. R. Kristensen, H. Madsen, and R. V. Overgaard. A matlab framework for estimation of NLME models using stochastic differential equations. *Journal of Pharmacokinetics and Pharmacodynamics*, 34(5):623–642, 2007.

[52] P. C. Mundinger. Animal cultures and a general theory of cultural evolution. *Ethology and Sociobiology*, 1:183 – 223, 1980.

[53] J. Nagumo, S. Arimoto, and S. Yoshizawa. An active pulse transmission line simulating nerve axon. *Proceedings of the IRE*, 50(10):2061–2070, 1962.

[54] S. Nowicki, D. Hasselquist, S. Bensch, and S. Peters. Nestling growth and song repertoire size in great reed warblers: evidence for song learning as an indicator mechanism in mate choice. *Proceedings of the Royal Society of London B: Biological Sciences*, 267(1460):2419–2424, 2000.

[55] R. V. Overgaard, N. Jonsson, C. W. Tornøe, and H. Madsen. Non-linear mixed-effects models with stochastic differential equations: Implementation of an estimation algorithm. *Journal of Pharmacokinetics and Pharmacodynamics*, 32(1):85–107, 2005.

[56] B. H. Patterson and L. A. Zech. Development of a model for selenite metabolism in humans. *The Journal of Nutrition*, 122(3S):709–714, 1992.

[57] A. R. Pedersen. A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scandinavian Journal of Statistics*, 22:55–71, 1995.

[58] K. B. Petersen and M. S. Pedersen. *The matrix cookbook*. Technical University of Denmark, 2008.

[59] P. C. Phillips and J. Yu. Maximum likelihood and Gaussian estimation of continuous time models in finance. In *Handbook of Financial Time Series*, pages 497–530. Springer, New York, 2009.

[60] U. Picchini and S. Ditlevsen. Practical estimation of high dimensional stochastic differential mixed-effects models. *Computational Statistics & Data Analysis*, 55(3):1426–1444, 2011.

[61] U. Picchini, A. De Gaetano, and S. Ditlevsen. Stochastic differential mixed-effects models. *Scandinavian Journal of Statistics*, 37(1):67–90, 2010.

[62] J. Pinheiro and D. Bates. *Mixed-effects models in S and S-PLUS*. Springer Science, New York, 2006.

[63] I. Potamitis, S. Ntalampiras, O. Jahn, and K. Riede. Automatic bird sound detection in long real-field recordings: Applications and tools. *Applied Acoustics*, 80:1–9, 2014.

[64] R. Poulsen. *Approximate maximum likelihood estimation of discretely observed diffusion processes*. CAF, Centre for Analytical Finance, University of Aarhus, 1999.

[65] B. Ribba, N. H. Holford, P. Magni, I. Trocóniz, I. Gueorguieva, P. Girard, C. Sarr, M. Elishmereni, C. Kloft, and L. E. Friberg. A review of mixed-effects models of tumor growth and effects of anticancer drug treatment used in population analysis. *CPT: Pharmacometrics & Systems Pharmacology*, 3(5):1–10, 2014.

[66] M. Sandsten, M. Große Ruse, and M. Jönsson. Robust feature representation for classification of bird song syllables. *EURASIP Journal on Advances in Signal Processing*, 2016:68, 2016.

[67] G. Sermaidis, O. Papaspiliopoulos, G. O. Roberts, A. Beskos, and P. Fearnhead. Markov chain monte carlo for exact inference for diffusions. *Scandinavian Journal of Statistics*, 40(2): 294–321, 2013.

[68] L. B. Sheiner and S. L. Beal. Evaluation of methods for estimating population pharmacokinetic parameters. i. michaelis-menten model: routine clinical pharmacokinetic data. *Journal of Pharmacokinetics and Biopharmaceutics*, 8(6):553–571, 1980.

[69] D. Slepian and H. O. Pollak. Prolate spheroidal wave functions, fourier analysis and uncertainty I. *Bell Labs Technical Journal*, 40(1):43–63, 1961.

[70] P. Somervuo, A. Härmä, and S. Fagerlund. Parametric representations of bird sounds for automatic species recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(6):2252–2263, 2006.

[71] H. Sørensen. Parametric inference for diffusion processes observed at discrete points in time: a survey. *International Statistical Review*, 72(3):337–354, 2004.

[72] D. Stucke, M. Große Ruse, and D. Lebelt. Measuring heart rate variability in horses to investigate the autonomic nervous system activity – Pros and cons of different methods. *Applied Animal Behaviour Science*, 166:1–10, 2015.

[73] O. Tchernichovski, T. J. Lints, S. Deregnaucourt, A. Cimenser, and P. P. Mitra. Studying the song development process: Rationale and methods. *Annals of the New York Academy of Sciences*, 1016(1):348–363, 2004.

[74] D. J. Thomson. Spectrum estimation and harmonic analysis. *Proc. of the IEEE*, 70(9): 1055–1096, Sept 1982.

[75] K. P. Timms, C. A. Martín, D. E. Rivera, E. B. Hekler, and W. Riley. Leveraging intensive longitudinal data to better understand health behaviors. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pages 6888–6891. IEEE, 2014.

[76] C. W. Tornøe, H. Agersø, E. N. Jonsson, H. Madsen, and H. A. Nielsen. Non-linear mixed-effects pharmacokinetic/pharmacodynamic modelling in NLME using differential equations. *Computer Methods and Programs in Biomedicine*, 76(1):31–40, 2004.

[77] C. W. Tornøe, R. V. Overgaard, H. Agersø, H. A. Nielsen, H. Madsen, and E. N. Jonsson. Stochastic differential equations in NONMEM: implementation, application, and comparison with ordinary differential equations. *Pharmaceutical Research*, 22(8):1247–1258, 2005.

[78] T. A. Walls and J. L. Schafer. *Models for intensive longitudinal data.* Oxford University Press, 2006.

[79] E. Węgrzyn and K. Leniowski. Syllable sharing and changes in syllable repertoire size and composition within and between years in the great reed warbler, Acrocephalus arundinaceus. *Journal of Ornithology*, 151(2):255–267, 2010.

[80] P. M. Woodward. *Probability and Information Theory with Applications to Radar.* McGraw-Hill, New York; Pergamon Press, London, 1953.

[81] H. Xie, R. E. Drake, S. J. Kim, and G. J. McHugo. Analyzing long-duration and high-frequency data using the time-varying effect model. *Administration and Policy in Mental Health and Mental Health Services Research*, 44(2):225–232, 2017.

[82] N. Yoshida. Estimation for diffusion processes from discrete observation. *Journal of Multivariate Analysis*, 41(2):220–242, 1992.