State Space Models for Regional Epidemiolgical Indicators

Stefan Heyder

October 2023 – Draft v 0.1

Abstract

Short summary of the contents in English...

Zusammenfassung

Kurze Zusammenfassung des Inhaltes in deutscher Sprache...

Publications and Contributions

This thesis consists of mostly unpublished work. During my time as a PhD student I have, however, been fortunate to collarborate with many scientists on problems in mathematical epidemiology with a focus on COVID-19, which resulted in several publications. In this section I want to clarify what my contributions to these publications were and which contributions of the present thesis are new.

Bracher, J. et al. (Aug. 27, 2021). "A Pre-Registered Short-Term Forecasting Study of COVID-19 in Germany and Poland during the Second Wave." In: *Nature Communications* 12.1 (1), p. 5173. ISSN: 2041-1723. DOI: 10.1038/s41467-021-25207-0.

Bracher, J. et al. (Oct. 31, 2022). "National and Subnational Short-Term Forecasting of COVID-19 in Germany and Poland during Early 2021." In: *Communications Medicine* 2.1 (1), pp. 1–17. ISSN: 2730-664X. DOI: 10.1038/s43856-022-00191-8.

Burgard, J. P. et al. (Aug. 31, 2021). "Regional Estimates of Reproduction Numbers with Application to COVID-19." arXiv: 2108.13842 [stat]. URL: http://arxiv.org/abs/2108.13842 (visited on 09/30/2021).

Grundel, S. et al. (Apr. 2022). "How Much Testing and Social Distancing Is Required to Control COVID-19? Some Insight Based on an Age-Differentiated Compartmental Model." In: *SIAM Journal on Control and Optimization* 60.2, S145–S169. ISSN: 0363-0129, 1095-7138. DOI: 10.1137/20M1377783.

Grundel, S. M. et al. (Jan. 1, 2021). "How to Coordinate Vaccination and Social Distancing to Mitigate SARS-CoV-2 Outbreaks." In: *SIAM Journal on Applied Dynamical Systems* 20.2, pp. 1135–1157. DOI: 10.1137/20M1387687.

Sherratt, K. et al. (June 16, 2022). *Predictive Performance of Multi-Model Ensemble Forecasts of COVID-19 across European Nations*. DOI: 10.1101/2022.06.16.22276024. Pre-published.

Du musst bereit sein Dinge zu tun.

— A meme on the internet, 2022.

Acknowledgments

Put your acknowledgments here.

Contents

Chapter 1

Introduction

The Coronavirus disease 2019 (COVID-19) pandemic put the scientific community to the test: how infectious, morbid and mortal was the dieseas? when and for how long did infected people become infectious? how effective are the countermeasures taken? how safe and effective are the vaccines that were developed at an unprecedented speed? Some of these questions, e.g. those about the epidemiology of COVID-19, are confined to well-established areas of research, while others, e.g. those about the efficacy of countermeasures, required collaboration across a wide range of disciplines: from infectious disease epidemiology, mathematical and statistical modeling, social and communication science to non-scientific actors such as legislators, journalists and politicians.

Although there is still a lot of scientific and societal follow-up work to be done, given the magnitude of this challenge, it is astonishing how well science and society as a whole have handled the pandemic. A key factor in this accomplishment is the large-scale availability of data surrounding the pandemic. In many countries, including Germany, data on reported cases, deaths, vaccinations and deaths were published daily by the respective national health authorities, i.e. the Robert Koch-Institut (RKI) (Robert Koch-Institut, 2021, 2022a) in Germany. Additionally, mobility data from mobile communications providers allowed researchers to relate human movement to the spread of COVID-19 (Kraemer et al., 2020; Schlosser et al., 2020). As each day the news reported on the number of newly reported cases and deaths, numerous dashboards with analyses of COVID-19 data were made available and an abundant number of scientific works was created, effectively communicating with the public, whose cooperation with countermeasures was critical, became more and more important. To disseminate insights to the public, we need to understand and communicate to them the underlying dynamics of an epidemic.

An epidemic outbreak is inherently a random phenomenon (Diekmann, Heesterbeek, and Britton, 2013). Who becomes infected, for how long they stay infectious, whom they meet while they are infectious and whom they finally infect are all aspects that depend to a certain degree on chance. If one is interested in large-scale phenomena, e.g. effects of immunization in a large population, one may get away with a deterministic model (Britton et al., 2019), such as the classical S(E)IR model (Kermack and McKendrick, 1927) or variants of it. However, as soon as one is interested in more detailed phenomena, as we are in this thesis, stochastic and statistical modeling becomes essential.

An epidemic outbreak is also inherently a local phenomenon, especially in the early phase of the epidemic. In the extreme case, there is only a single infectious person and, for the most part, their potential infectees will belong to the same spatial region as the infector. Therefore, we should incorporate this locality into our models. To fit such models to data, the data has to include spatial information. Luckily, the case and death data are available at the subnational level in most countries. In Germany, it is even available at the county (Landkreis, NUTS3) level (Robert Koch-Institut, 2022a).

As statisticians, having access to such a large amount of data is both a blessing and a curse. While more, and ideally better, data allows us to formulate and answer more relevant questions, the models we create to accommodate these data become more and more intricate. Intricate models require

more care in modeling, fitting and interpretation, as more things can go wrong along the way. Thus we will tread carefully. As we incorporate more detailed effects into our models, fitting the models to data becomes difficult to practically impossible using established techniques. While there are some remedies for this curse of dimensionality, e.g. exploiting as much available structure as possible, there is an ongoing need for new procedures enabling inference in these settings. Additionally, we need mathematical as well as practical insight into the performance of these procedures to make informed decisions in applied settings: which methods should we prefer under which circumstances?

These considerations set the stage for this thesis. Driven by the need for good statistical models that allow us to answer urgent questions in infectious disease epidemiology, with COVID-19 as a driving example, we will start with an analysis of what is required of these sought-after models. We will define and discuss the role of several epidemiological indicators, i.e. quantities that have an interpretation related to the epidemic. It turns out that we will usually be interested in quantifying the speed at which the epidemic proliferates, and we discuss several popular indicators that measure this speed. A useful statistical analysis should provide interpretable insight into the problem at hand, so we focus on how straightforward this interpretation is, giving recommendations on when to use which indicator. To estimate these indicators from data, we have to create statistical models that include them. Before we do so, we will create a list of desiderata from the context of COVID-19.

Once we have a clear view of the epidemiological problems at hand, we show that many of the desiderata can be covered by using SSMs, a flexible framework for modeling non-stationary time series. Unfortunately, we will require that these SSMs include integer-valued, non-Gaussian, observations, which makes fitting the models to data analytically impossible and numerically difficult, as one is essentially faced with a high-dimensional non-Gaussian Bayesian inference problem. Instead, inference will be based on simulation methods, most notably importance sampling. To apply these methods, the practitioner has some flexibility in the so-called proposal distribution, a tractable approximation to the Bayesian posterior. Different disciplines have developed simulation-based techniques that allow the user to choose optimal proposals, where optimality is based on different performance criteria for different methods. In this thesis, we focus on two methods: the Cross-Entropy method (CE-method) and Efficient Importance Sampling (EIS). In the literature, a comparison between these two methods is missing: there are neither mathematical nor empirical results comparing the two. We fill this gap by first proving central limit theorems for both methods, allowing for a theoretical comparison. Additionally, we also provide simulation studies comparing the methods on instructive univariate and SSMs examples. To this end, we also develop a new algorithm that allows the CE-method to be applied to state space models (SSMs).

Finally, we demonstrate how to solve a selection of infectious disease epidemiology problems using the mathematical insights we gained. These examples focus on the COVID-19 epidemic in Germany and illustrate the modeling, computational and applied aspects of this thesis.

add some more refs?

Chapter 2

Epidemiological considerations

Contributions of this chapter

Objectives of epidemiological modelling
Available data and its quality
Desiderata for epidemiological models

The spread of infectious diseases, such as COVID-19, is a complex phenomenon. For COVID-19, this complexity arises from the interplay of many factors. Studying these influences allows us to define the aims and challenges of epidemiological modeling in the context of this thesis. It also will guide us towards desirable and possible outcomes of our efforts from an applied perspective.

First of all, there is considerable heterogeneity in the way the disease progresses once an individual is infected (Salzberger et al., 2021). Some infectees may show few to no symptoms but are still highly infectious (Byambasuren et al., 2020), and disease progression is tightly linked to age and preexisting comorbidities (Biswas et al., 2020). Additionally, different variants of SARS-CoV-2 differ in key epidemiological characteristics such as the reproduction number (Du et al., 2022) and mortality (Hughes et al., 2023).

Second, the spread is highly dependent on the contact behavior in the population, as the infector has to be in close physical proximity to the infectee to infect them. These contact patterns are an essential component of any mathematical model for infectious diseases, as they define how the epidemic evolves. While there are some empirical studies (Mossong et al., 2008; Tomori et al., 2021), capturing the contact behavior at certain points in time, in the context of an ongoing epidemic these patterns are subject to change, not only in intensity but also in shape (Tomori et al., 2021). As contact restrictions were put into place or lifted, mask-wearing was enforced and home office and schooling became commonplace, so did the number of contacts change and occur under different circumstances.

Finally, as the virus spread in the population and vaccinations became available, the population became partially immune against the disease, if not against infection

citep

. This immunity affects the spread as well: if an infector has contact with a partially immune individual, the probability of transmission is smaller. Additionally, partial immunity may lead infectors to develop fewer or no symptoms so they may not be aware of being infectious, foregoing quarantine.

Parts of this chapter, especially Sections 2.2 to 2.4, consist of the ideas published in (Heyder and Hotz, 2023), but have been rewritten to fit better into this thesis.

As statisticians, we are faced with a difficult problem: Which of these factors should we include in our model and how? The answer certainly depends on the epidemiological question under consideration and the availability and quality of data.

2.1 Objectives of epidemiological modelling

Before considering the mathematical modeling of epidemics, let us make clear what the goals of our investigation are. In this thesis, we are interested in providing models that are informed by real-world data, allow us to learn about the past, current or future state of the epidemic and whose results are, ideally, easy to communicate to non-experts, e.g. political stakeholders. These time scales can be translated into the following three tasks for epidemiological modeling.

Retrospective Analysis Here we are interested in an ex-post analysis of a period of interest in the past. The goal here is either to infer intrinsic epidemiological quantities, such as the time-varying reproduction number R_t (Abbott et al., 2020) or to evaluate the performance of non-pharmaceutical interventions (NPIs) taken (Brauner et al., 2021; Flaxman et al., 2020; Khazaei et al., 2023). The results of this analysis may inform future decisions on which countermeasures to implement, and as such we want a causal link between the NPIs prescribed and the reduction in reported cases. Naturally, this is a difficult objective to accomplish due to several aspects. The data at our disposal is observational and there are several quality issues, see Section 2.3. Additionally, the interplay between NPIs and change in the behavior of the population is intricate, where voluntary behavioral change may precede the enforced social distancing (Gupta, Simon, and Wing, 2020). For some examples focusing on the efficacy of NPIs, we refer the reader to the excellent articles (Brauner

et al., 2021; Flaxman et al., 2020; Khazaei et al., 2023), especially the discussion and limitation sections therein. In these types of analyses, we can assume that all data related to that period is as complete as it will be. Methods used to perform these analyses range from estimating parameters for each day individually, e.g. using the EpiEstim (Cori, 2021) method (Abbott et al., 2020), to constructing complex Bayesian mechanistic (Flaxman et al., 2020) and hierarchical models (Brauner et al., 2021; Khazaei et al., 2023).

Monitoring For monitoring, we are interested in real-time inference about the current state of the epidemic. This includes the recent past and near future and may include now- and forecasts of cases, hospitalizations or deaths. Here data is not yet final, and inference is complicated by slow reporting and data revisions, see Section 2.3. The results of monitoring can be used to inform current policy, i.e. whether current NPIs should be lifted or new ones enforced. Most online dashboards that emerged at the beginning of the pandemic fall into this category. The result of monitoring may either be an estimate of an epidemiological indicator, but may also consist of short-term forecasts. Examples of the former include the daily reproduction number estimates of the RKI (An Der Heiden and Hamouda, 2020), the Helmholtz Centre for Infection Research's dashboard (Khailaie et al., 2021) or the dashboard of the authors team (Hotz et al., 2020).

While some of these dashboards also provide forecasts of cases, a more concerted effort of forecasts is provided by the U.S. ForecastHub (Ray et al., 2020), its German/Polish (Bracher et al., 2021; Bracher et al., 2022) and EU/EFTA (Sherratt et al., 2022) equivalents. These collaborative platforms gathered real-time forecasts of COVID-19 cases and deaths in the upcoming four weeks, based on an ensemble that aggregates predictions from several models provided by expert modelers. In a real-time setting, these forecasts can be evaluated which may inform practitioners as to which model to prefer. For forecasting, methods range from classical time series analysis methods (Arroyo-Marioli et al., 2021) to compartmental models (Khailaie et al., 2021) and computationally intensive agent-based models (Adamik et al., 2020).

Scenario Modeling Scenario modeling concerns itself with the impact that changes of current circumstances, e.g. variants, seasonality, policies, vaccination or NPIs, have on public health outcomes. Contrary to monitoring, the goal is to quantify the influence over longer periods with scenarios reaching multiple months into the future. The parameters of scenarios are assumed to be uncertain as well, making the task at hand challenging. These forecasts are difficult to evaluate, as the scenario specifications rely on assumptions that are hard to verify in practice. Nevertheless, these scenarios help policymakers make informed decisions (Borchering et al., 2023).

In the context of this thesis, we are primarily interested in performing retrospective analyses and providing tools for monitoring as well as short-term forecasting. While scenario modeling has its own merits, evaluating the performance of models is much harder, as there is no ground truth to compare against. Additionally, the methods developed in this thesis rely on having recurring observations on a daily or weekly time scale, usually in the form of reported cases, deaths, or hospitalizations, which for scenario modeling are not available. If such observations are not available, i.e. because we are forecasting months ahead, the uncertainty produced by our models will be much too large to be sensible.

2.2 Measures of epidemic spread

A key component of any epidemiological model is how the spread of the epidemic is accounted for. As argued before, an epidemic is a complex process, driven by many different factors. To make this complexity manageable we employ simplified models for the spread of cases and, depending on the assumptions made, different measures that quantify the epidemic's spread arise. Actually, we are not only interested in the spread of the epidemic but also in the speed, i.e. the change over time,

with which cases proliferate, because it allows us to make predictions about future cases and thus give recommendations about whether countermeasures should be employed or lifted.

In this thesis, we will primarily focus on two of these measures: the growth factor and the reproduction number. As we will argue, these two measures come with simple interpretations and, as such, are valuable in communicating the results of our modeling efforts not only to other researchers but also to non-experts such as the public and political stakeholders.

Additionally, we will be interested in measures that capture the severity of the epidemic, i.e. the morbidities and mortalities caused by the epidemic. As these events are consequences of infection that occur after a delay, they can be recovered from incidence data. Thus modeling the spread of the epidemic serves two goals: making inferences and predictions about the cases and associated measures, as well as morbidities.

To introduce the different measures in the following, we will, for the moment, make some simplifying assumptions about the population in which the epidemic spreads and the time frame considered. Consecutively, we will relax these assumptions to accommodate more realistic populations.

First of all, we consider a homogenous population with homogenous mixing. This means that any two individuals in the population are affected by the epidemic in the same way: the probabilities of becoming infected, infectious, hospitalized or recovering from infection are the same for every individual in the population. Additionally, homogenous mixing indicates that once an individual is infected, they meet and infect every other individual in the population with the same probability. Furthermore, we assume that the population is large enough that the probability of duplicate infections, i.e. becoming infected twice either from the same or different individuals, is negligibly small, and we assume that infections occur independently from one another. Similarly, we could also assume that the population is infinitely large or that the time frame under consideration is sufficiently short. Finally, we assume that the behavior of the population is constant over the period modeled.

2.2.1 Reproduction number

We will model the evolution of the epidemic in discrete time, as this is the time scale on which data are available. Denote by $I_0 \in \mathbf{N}$ the initial number of infected and for a day $t \in \mathbf{N}$ let I_t be the number of newly infected individuals on that day. Note that I_t is random. For $\tau \in \mathbf{N}$ let β_τ be the expected number of secondary cases a primary case infects τ days after they become infectious themselves and assume that the expected number of secondary cases $R_c = \sum_{\tau \in \mathbf{N}} \beta_\tau$ is non-zero and finite $0 < R_c < \infty$. R_c is called the case reproduction number. Here we have implicitly assumed that $w_0 = 0$, i.e. infected individuals need at least one day to become infectious themselves. For COVID-19 this is a reasonable assumption (Lauer et al., 2020).

As we have assumed that R_c is finite, we may write $\beta_\tau = R_c w_\tau$ where $w_\tau = \frac{\beta_\tau}{R_c}$. $w = (w_\tau)_{\tau \in \mathbf{N}}$ is called the generation time distribution or the infectivity profile. On day t the conditional expectation of newly infected individuals given all past incidences $\mathbb{E}\left(I_t | I_{t-1}, I_{t-2}, \dots\right)$ can then be written as a convolution of w and the number of past cases

$$\mathbb{E}\left(I_t | I_{t-1}, I_{t-2}, \dots\right) = R_c \sum_{\tau=1}^{\infty} I_{t-\tau} w_\tau, \tag{2.1}$$

the so-called renewal equation. Here $I_{t-\tau} = 0$ if $\tau > t$. If case numbers are small, e.g. if the assumptions we demand hold, the conditional distribution of I_t given past cases is, by the law of small numbers, well approximated by a Poisson distribution and combined with the renewal equation we obtain the renewal equation model (C. Fraser, 2007)

$$I_t | I_{t-1}, I_{t-2}, \dots \sim \text{Pois}\left(R_t \sum_{\tau=1}^{\infty} I_{t-\tau w_\tau}\right), \tag{2.2}$$

where the time-varying or instantaneous reproduction number R_t is now allowed to vary over time as well. Working with the time-varying reproduction number R_t over the case reproduction number R_c

has the advantage that R_t can be estimated from data until day t alone, while $R_c = \sum_{\tau=1} w_\tau R_{t+\tau}$ depends on future cases (C. Fraser, 2007).

Given incidence data $I_t, I_{t-1}, \dots$ we can perform frequentist inference on R_t in Equation (2.2) by estimating

$$\hat{R}_t = \frac{I_t}{\sum_{\tau=1}^{\infty} I_{t-\tau} w_\tau},$$

which is a moment- and maximum-likelihood estimator (Hotz et al., 2020). Additionally (Cori, 2021) provides a Bayesian framework, using conjugate gamma priors for R_t. If one is interested in the case reproduction number R_c it can be recovered from estimates of R_t as $\hat{R}_c = \sum_{\tau=1}^{\infty} w_\tau \hat{R}_{t+\tau}$ or using the Wallinga-Teunis estimator (Wallinga and Teunis, 2004).

The reproduction numbers R_c and R_t have a mechanistic interpretation: R_c is the number of secondary cases an infectious individual can expect to infect over the time of their disease, and so is R_t with the additional assumption that the behavior of the infection process stays the same for the whole duration of infection. Assuming that contacts lead to infection independently and with the same probability, this means that the reproduction numbers are proportional to the total number of contacts a person has, and as such, they are an excellent measure of the efficacy of NPIs (Brauner et al., 2021; Flaxman et al., 2020; Khazaei et al., 2023). An additional advantage is that the model (2.2) can be interpreted mechanistically, i.e. R_t gives a mechanical model of why the number of cases increases. In contrast, the model of exponential growth that we will address next is more phenomenological in nature, based on the observation that the number of cases tend to increase or decrease exponentially.

2.2.2 Growth Factor

If I_0 is small compared to the total population size, a sensible assumption, one can show that under the above model, the expected number of cases grows approximately exponentially (Diekmann, Heesterbeek, and Britton, 2013, Section 1.2),

$$\mathbb{E}I_t \approx \rho \mathbb{E}I_{t-1} \approx \rho^t \mathbb{E}I_0. \tag{2.3}$$

ρ is called the daily exponential growth factor and can be recovered from Equation (2.1) by an exponential ansatz (Wallinga and Lipsitch, 2007):

$$\rho^t \mathbb{E}I_0 = \mathbb{E}I_t = R_c \sum_{\tau=1}^{\infty} I_{t-\tau} w_\tau = R_c \sum_{\tau=1}^{\infty} \mathbb{E}I_0 \rho^{t-\tau} w_\tau = \rho^t \mathbb{E}I_0 \left(R_c \sum_{\tau=1}^{\infty} \rho^{-\tau} w_\tau \right)$$

which shows that unless ρ or $\mathbb{E}I_0$ is zero,

$$\sum_{\tau=1}^{\infty} \rho^{-\tau} w_\tau = \frac{1}{R_c}$$

has to hold. The left-hand side is the probability generating function $\mathbb{E}\rho^{-W}$ for $W \sim \sum_{\tau=1}^{\infty} w_\tau \delta_\tau$. As $W \geq 1$ almost surely and the probability generating function is strictly increasing with limits 0 and ∞ as ρ goes to 0 and ∞ respectively , there is exactly one solution $\rho \in \mathbf{R}_{>0}$ to this equation. Here we assume that w is not degenerate, i.e. not a.s. 1 and that we can exchange limits with the infinite sum, e.g. because w has bounded support. Thus, once the infectivity profile w is fixed, there is a one-to one relationship between ρ and R_c.

Similarly to the time-varying reproduction number, we may alter Equation (2.3) by introducing for $t \in \mathbf{N}$ a time-varying growth factor $\rho_t \in \mathbf{R}_{>0}$, resulting in

$$\mathbb{E}I_t = \rho_t \mathbb{E}I_{t-1},$$

which can be estimated, e.g. by the moment-estimator $\hat{\rho}_t = \frac{I_{tj}}{I_{t-1}}$.

Focusing on the growth factor over the reproduction number has the advantage that one does not need to specify a generation time distribution w to estimate ρ_t, whereas it is essential for estimating R_t.

2.2.3 Other indicators

Instead of concentrating on the daily evolution of the epidemic, it may be beneficial to consider the weekly behavior instead. As we will see in Section 2.3, the incidence data available in Germany are strongly contaminated by weekday effects, with few cases reported on the weekends and more during the week. To avoid explicitly modeling these effects we will, in Section 4.2 group the case data by weeks and estimate the weekly growth factor

$$\rho^7 \approx \frac{\mathbb{E}\sum_{s=0}^{6} I_{t-s}}{\mathbb{E}\sum_{s=0}^{6} I_{t-7-s}}.$$

Here we assumed, again, that the circumstances of the epidemic do not change over the period considered. By slight abuse of notation, we let ρ_t^7 be this weekly growth factor, where now t is counting weeks instead of days. Notice that when ρ_t is time-varying, it is not necessarily the case that $\rho_t^7 = \prod_{s=0}^{6} \rho_{t-s}$. Notice that if $\mathbb{E}W = 7$, i.e. the average infectious period is one week, the delta-method yields

$$R_c = \frac{1}{\mathbb{E}\rho^{-W}} \approx \rho^7.$$

Let us hasten to add that there is no reason for the error in this approximation to be small. Nevertheless, as the average infectious period is somewhat smaller than one week for COVID-19, we may think of ρ^7 as, approximately, the reproduction number.

The exponential growth rate r and doubling time d are closely related to the growth factor, and are given by

$$r = \log \rho \qquad\qquad d = \log_\rho 2 = \frac{\log 2}{\log \rho}.$$

Thus r is the growth rate of the exponentially increasing cases, $\mathbb{E}I_t \approx \exp(rt)\mathbb{E}I_0$ and d is the time it takes for cases to double under this exponential growth, $\mathbb{E}I_{t+d} \approx \mathbb{E}2I_t$. Notice that the last equation only makes sense if $d \in \mathbf{N}$, or if we model the epidemic to evolve in continuous time instead. These two quantities are not as easy to interpret as, e.g., the weekly growth factor or the reproduction number.

2.2.4 Usefulness of indicators for communication

This subsection briefly summarizes the ideas published in (Heyder and Hotz, 2023).

Given incidence data, which of the above indicators should one estimate and report? The answer depends, of course, on the goals of one's investigations and the data at hand as well as the audience to which one communicates these estimates.

When the audience is the general public, reproduction numbers and growth factors allow us to convey the exponential growth of the epidemic, either by an argument based on generations (reproduction number) or by exponential growth in time (growth factor).

If data about the infectivity profile w is available, e.g. from contact-tracing studies, reproduction numbers have the advantage of having a concrete, mechanistic interpretation as numbers of infectious contacts. If one is interested in containing the epidemic, i.e. „flattening the curve“, reducing contacts uniformly in the population by a factor of $c = 1 - \frac{1}{R}$ works. That is if $R = 1.25$, we have to reduce contacts by 20% to reach $R = 1$. Such information is useful not only for policymakers but also for the general population.

Another concern of monitoring is short-term forecasts of future cases, say, for one to four weeks ahead. Given just the estimated growth factor or reproduction number, this task is more easily achieved by the growth factor: it suffices to multiply current incidences by ρ^7 to get the expected number of cases in the next week. For reproduction numbers, forecasting is more involved, relying on simulation to repeatedly sample from Equation (2.2).

We recommend not communicating exponential growth rates and doubling times if at all possible. Exponential growth rates come with the disadvantages of the growth factor without the upside of

having easily accessible forecasts. While doubling times allow for forecasts in terms of when the number of cases will double, such forecasts are usually not of primary interest.

How well we can estimate these indicators depends on the available data and, in particular, their quality.

2.3 Available data and its quality

- surprising amount of data available, but quality questionable,
- in Germany have data on reported cases and deaths by gender, age group, county, with reporting date of case and for some cases even date of symptom onset
- reporting of cases is regulated by Infektionsschutzgesetz
- parallel dataset for reports of hospitalisations
- have description section from Nowcasting draft here
- descriptive statistics of German COVID-19 data set
- even larger datasets that compile this for europe + EFTA (?) by ECDC or by world (JHU)
- quality of reported case data is potentially too low
  - reporting delays
  - weeakday effects
  - testing regime changing (2G/3G)
  - ...
- data on commuting

2.4 Desiderata for epidemiological models

As the dynamics of the epidemic are constantly changing, our models should

- we want models to be able to include as much data as possible, while still being numerically tractable

this paragraph to modelling chapter

The Poisson distribution arises from the law of small numbers: if there is a large population where every individual has, independently, a small probability of becoming infected in a small window of time then the total number of infections in that window of time is well approximated by the Poisson distribution. Indeed, the law of small numbers remains valid for small dependencies Arratia, Goldstein, and Gordon, 1990; Ross, 2011. However, incidences observed from the SARS-CoV-2 epidemic tend to follow a negative binomial distribution S. Chan et al., 2021.

Regional dependencies and effects

- German case data are reported on Landkreis level, performing analysis of each individual is not sensible
- inhabitants travel between regions, and measures were taken on on regional level as well
- effects are not really spatial: euclidean distance is not so much of an issue but how closely connected regions are (give some examples)
- also want to account for other regional effects such as different socio-economic settings ...

Temporal correlation

Interpretability

Chapter 3

Importance Sampling in State Space Models

Contributions

The main contribution of this chapter consists of a rigorous comparison of two importance sampling frameworks: the Cross-Entropy method (CE-method) and Efficient Importance Sampling (EIS). Both methods determine optimal importance sampling proposals, but have, until now, been studied in separate communities: the CE-method is popular in rare-event estimation and engineering disciplines, while EIS is popular in the financial time series community.
The contributions of the individual sections are as follows:

Modeling epidemiological desiderata with state space models

Gaussian Linear State Space Models This section is a condensed introduction to Gaussian linear state space models (GLSSMs) and is loosely based on (Durbin and Koopman, 2012).

Partially Gaussian state space models

Importance Sampling

- We prove Lemma 3.5.
- Discussion surrounding (Chatterjee and Diaconis, 2018).
- We prove central limit theorems for both methods (Sections 3.4.2 and 3.4.3).
- Proof Proposition 3.5.

Interim discussion

Gaussian importance sampling for state space models derive an efficient algorithm to apply the CE-method to state space models (Section 3.6.2)

Maximum likelihood estimation in SSMs

Comparison of Importance Sampling method Extensively compare both methods on theoretical as well as practically relevant properties with instructive univariate and multivariate examples (Section 3.8).

State space models (SSMs) form a versatile class of statistical models that allow to modeling of non-stationary time series data while providing a straightforward, mechanistic interpretation of the time series' dynamics. The main idea of these models is to introduce unobserved **latent states** whose joint distribution is given by a Markov process and model the observed time series conditional on these states. By exploiting this structure, inference in SSMs becomes computationally efficient, i.e. the complexity of algorithms is usually linear in the number n of time points considered. In this chapter, we provide a mathematical introduction to the theory of SSMs and the main tool we will use for inference, importance sampling. Additionally, we will highlight how to use SSMs to model the desiderata identified in Section 2.4.

Let us begin with the most general definition of a SSM.

Definition 3.1 (State Space Model)**.** A **SSM** is a discrete time stochastic process $(X_t, Y_t)_{t=0,\dots,n}$ taking values in the measurable space $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}_\mathcal{X} \otimes \mathcal{B}_\mathcal{Y})$ such that

(i) The marginal distribution of the **states** $(X_0, \dots, X_n)$ is a discrete time Markov process, i.e. for $t = 1, \dots, n$

$$\mathbf{P}\left(X_t \in B | X_0, \dots, X_{t-1}\right) = \mathbf{P}\left(X_t \in B | X_{t-1}\right) \text{ a.s.} \tag{3.1}$$

for all measurable $B \in \mathcal{B}_\mathcal{Y}$.

(ii) Conditional on the state X_t and observation Y_{t-1}, Y_t is independent of X_s and Y_{s-1}, $s < t$, i.e.

$$\mathbf{P}\left(Y_t \in B | X_0, \dots, X_t, Y_0, \dots, Y_{t-1}\right) = \mathbf{P}\left(Y_t \in B | X_t, Y_{t-1}\right)$$

for all measurable $B \in \mathcal{B}_\mathcal{Y}$.

For notational convenience, we will write $X_{s:t} = (X_s, \dots, X_t)$ for the vector that contains all states from s to t, $s \leq t$, dropping the first index if we consider the whole set of observations up to time t, so $X_{:t} = X_{0:t}$, and dropping the subscript if we consider all states at once, $X = X_{:n}$. Similarly we set $Y_{s:t} = (Y_s, \dots, Y_t)$, $Y_{:t} = Y_{0:t}$ and $Y = Y_{:n}$.

picture of dependency structure

The models that we consider in this thesis will usually admit densities for the state transitions w.r.t. a common dominating measure $\mu_\mathcal{X}$ and similar for the observations w.r.t. some dominating measure $\mu_\mathcal{Y}$.

check whether models in Ch4 violate this

Notation 3.1 (Densities, conditional densities)**.** We will use the standard abuse of notation for densities that makes the type of density „obvious“ from the arguments used. This means that $p(x)$ is the density for all states X, $p(x_t|x_{t-1})$ the conditional density of $X_t|X_{t-1}$ and similarly for observations: $p(y|x)$ is the density of all observations Y conditional on all states X.

Note that this notation also implicitly includes the time t and allows for changes in, e.g., the state transition over time.

When densities come from a parametric model parametrized by $\theta \in \Theta \subseteq \mathbf{R}^l$ and the dependence of the model on θ is of interest, i.e. because we try to estimate θ, we indicate this by adding a subscript to the densities. If the dependence is not of interest, e.g. because θ is fixed, I will usually omit θ for better readability.

In this notation, the joint density of a parametric SSM factorizes as

$$\begin{aligned} p_\theta(x, y) &= p_\theta(x_0, \dots, x_n, y_0, \dots, y_n) \\ &= p_\theta(x_0) \prod_{t=1}^{n} p_\theta(x_t|x_{t-1}) \prod_{t=0}^{n} p_\theta(y_t|x_t, y_{t-1}), \end{aligned} \tag{3.2}$$

where $p_\theta(y_0|x_0, y_{-1}) = p_\theta(y_0|x_0)$.

As inferences we make in this thesis depend on the SSM only through the likelihood we identify almost sure versions of (X, Y) with itself, i.e. all equations involving X or Y are understood almost surely.

Remark 3.1 (dependence on Y_{t-1}, dimensions)**.** Contrary to the standard definition of a SSM, our Definition 3.1 allows Y_t to depend on Y_{t-1}. As the models considered in Chapter 4 will make extensive use of SSMs with this dependency structure we opt to use this non-standard definition here. This is not a limitation of the standard definition: given a SSM of the form in Definition 3.1 we can transform it to the standard form by choosing states $(X_t, Y_t) \in \mathcal{X} \times \mathcal{Y}$ and observations $Y_t \in \mathcal{Y}$ such that the SSM becomes a stochastic process on $(\mathcal{X} \times \mathcal{Y}) \times \mathcal{Y}$.

Additionally, the goal of our inferences will always be the conditional distribution $X|Y$ for a single, fixed, set of observations Y. Assuming all densities exist, the conditional density $p(x|y)$ is given, up to a constant not depending on x, by Equation (3.2):

$$p(x|y) \propto p(x, y) = p(x_0) \prod_{t=1}^{n} p(x_t|x_{t-1}) \prod_{t=0}^{n} p(y_t|x_t, y_{t-1}).$$

Thus, the dependence of Y_t on Y_{t-1} only affects our inferences through $p(y_t|x_t, y_{t-1})$, where, as Y_{t-1} is observed, the argument y_{t-1} is fixed. Consequently, all results we present in this chapter for SSMs where Y_t depends only on X_t that concern only the conditional distribution $X|Y = y$ carry over to those given by Definition 3.1.

In most SSMs we consider in this thesis we use $\mathcal{X} = \mathbf{R}^m$, $\mathcal{Y} = \mathbf{R}^p$ or $\mathcal{Y} = \mathbf{Z}^p$ so that $\mathcal{X}$ is m dimensional and $\mathcal{Y}$ is p dimensional and equip these spaces with the usual σ-Algebras. Unless noted otherwise, we use for $\mu_{\mathcal{X}}$ the m-dimensional Lebesgue measure and for $\mu_{\mathcal{Y}}$ either the p-dimensional Lebesgue measure ($\mathcal{Y} = \mathbf{R}^p$) or the p-dimensional counting measure ($\mathcal{Y} = \mathbf{Z}^p$).

Given data $(y_t)_{t=0,\dots,n-1}$ that may be modeled with a SSM the practitioner is confronted with several tasks, which provide the structure of this chapter:

(i) Choosing a suitable, usually parametric, class of SSMs that include the effects of interest.

(ii) Fitting such a parametric model to the data at hand by either frequentist or Bayesian techniques.

(iii) Infer about the latent states X from the observations Y by determining, either analytically or through simulation, the smoothing distribution $X|Y$.

The first step, Item (i), requires that the practitioner specifies a joint probability distribution for the states and observations (Section 3.1). Due to the assumed dependency structure, this boils down to specifying transition kernels for the states and observations. The setting Definition 3.1 is too abstract to perform inference in, so further assumptions on the types of distributions for the latent states and observations are needed. In this chapter, we will discuss Gaussian linear state space model (GLSSM) (Section 3.2), where both the posterior distribution and the likelihood are analytically available. For the epidemiological application we have in mind these are however insufficient due to the non-linear behavior of incidences and the low count per region (Section 2.4). Such observations are better modeled with distributions on the natural numbers, i.e. with a Poisson or negative binomial distribution, both of which are exponential families of distributions. This will lead to the class of Logconcave state space models (LCSSMs) (Section 3.3) which will become the main focus of our study.

Regarding the second step, Item (ii), a frequentist practitioner will want to perform maximum likelihood inference on θ. While asymptotic confidence intervals for the maximum likelihood estimator (MLE) $\hat{\theta}$ can be derived both theoretically and practically (Durbin and Koopman, 2012, Chapter 7), they are, in the context of this thesis, usually of little interest. For these asymptotic frequentist procedures to be meaningful, an appropriate central limit theorem has to hold. However, as the time series we study are non-stationary and the dependence on parameters θ is allowed to be arbitrary, it is in general not obvious that such a theorem holds for the model under consideration. Instead, we choose to view this fitting as an Empirical Bayes procedure and our main practical interest lies in analyzing the posterior distribution $X|Y$ where we set θ equal to $\hat{\theta}$.

To obtain the maximum likelihood estimates $\hat{\theta}$ one needs access to the likelihood

$$p(y) = \int_{\mathcal{X}^n} p(x, y)\,\mathrm{d}x = \int p(y|x)p(x)\,\mathrm{d}x \tag{3.3}$$

which is usually not analytically available. Direct numerical evaluation of Equation (3.3) is hopeless due to the high dimensionality of the state space $\mathcal{X}^n$. Instead, we will resort to simulation-based inference by importance sampling (see Section 3.4), a Monte-Carlo method that approximates $p(y)$ by constructing a global tractable approximation to the integrand in Equation (3.3). Alternatively, sequential Monte Carlo (SMC) methods, i.e. particle filters, that perform importance sampling sequentially across the $n+1$ time steps can be used. We will not follow this approach for reasons described further below, but refer the reader to the excellent reference (Chopin and Papaspiliopoulos, 2020) for an introduction to these methods.

The performance of these simulations depends crucially on our ability to construct distributions that are close to the posterior $p(x|y)$ but are easy to sample from. To this end, we construct either Gaussian linear state space models (GLSSMs) (Section 3.6.1) in which sampling from the posterior is analytically possible, or Gaussian Markov processes (Section 3.6.2) which are directly amenable to simulation.

As an alternative to the MLE approach, a fully Bayesian approach would regard θ as random and administer a prior distribution, say with density $p(\theta)$. In this setting, the main interest still lies in determining the posterior distribution of $X|Y = y$, but due to the prior put on θ, its density, should it exist, is now given by

$$p(x|y) = \int p(x, \theta|y)\,\mathrm{d}\theta,$$

where $p(x, \theta|y)$ is the joint posterior of states and hyperparameters, conditional on observations y. To tackle this problem, one may again use importance sampling methods, see e.g. (Durbin and Koopman, 2012, Chapter 13.1), or use MCMC-methods tailored to SSMs, e.g. Particle-MCMC (Chopin and Papaspiliopoulos, 2020, Chapter 16).

3.1 Modeling epidemiological desiderata with state space models

3.2 Gaussian Linear State Space Models

Gaussian linear state space models (GLSSMs) are the working horses of most methods used in this thesis because they are analytically tractable and computationally efficient. Indeed for fixed dimension of states m and observations p the runtime of algorithms that we consider in this thesis is $\mathcal{O}(n)$.

Definition 3.2 (GLSSM)**.** A Gaussian linear state space model (GLSSM) is a joint distribution over states and observations (X, Y) where states a.s. obey the transition equation

$$X_{t+1} = A_t X_t + u_t + \varepsilon_{t+1} \qquad t = 0, \dots, n-1, \tag{3.4}$$

and observations a.s. obey the observation equation

$$Y_t = B_t X_t + v_t + \eta_t \qquad t = 0, \dots, n. \tag{3.5}$$

Here $A_t \in \mathbf{R}^{m\times m}$ and $B_t \in \mathbf{R}^{p\times m}$ are matrices that specify the systems dynamics. The **innovations** $(\varepsilon_{t+1})_{t=0,\dots,n-1}$ and **measurement noise** $(\eta_t)_{t=0,\dots,n}$ and the starting value $X_0 \sim \mathcal{N}(\mathbb{E}X_0, \Sigma_0)$ are jointly independent. Furthermore, $\varepsilon_{t+1} \sim \mathcal{N}(0, \Sigma_t)$ and $\eta_t \sim \mathcal{N}(0, \Omega_t)$ are centered Gaussian random variables and $u_t \in \mathbf{R}^m, t = 0, \dots, n-1$, $v_t \in \mathbf{R}^p, t = 0, \dots, n$ are deterministic biases.

Remark 3.2. From Equation (3.4) it is easy to see that the states $X = (X_0, \dots, X_n)$ form a Gaussian Markov process and that conditional on X_t, $t \in \{0, \dots, n\}$, Y_t is independent of X_s and Y_s, $s < t$. Thus A GLSSM is indeed a SSM.

The defining feature of a GLSSM is that the joint distribution of (X, Y) is Gaussian, as (X, Y) may be written as an affine combination of the jointly Gaussian $(X_0, \varepsilon_1, \ldots, \varepsilon_n, \eta_0, \ldots, \eta_n)$ and it is often useful to perform inferences in terms of innovations and measurement noise instead of states, see e.g. (Durbin and Koopman, 2012, Section 4.5).

As the joint distribution of (X, Y) is Gaussian, so are conditional distributions of states given any set of observations.

Lemma 3.1 (Gaussian conditional distributions)**.** *Let (X, Y) be jointly Gaussian with distribution $\mathcal{N}(\mu, \Sigma)$ where*

$$\mu = (\mu_X, \mu_Y)$$

and

$$\Sigma = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix},$$

where μ and Σ are partitioned according to the dimensions of X and Y.

Then the following holds:

(i) If Σ_{YY} is non-singular, $X|Y = y$ follows a Gaussian distribution with conditional expectation

$$\mu_{X|Y=y} = \mathbb{E}(X|Y = y) = \mu_X + \Sigma_{XY}\Sigma_{YY}^{-1}(y - \mu_Y)$$

and conditional covariance matrix

$$\Sigma_{X|Y=y} = \mathrm{Cov}(X|Y = y) = \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}.$$

(ii) In particular, let $X \sim \mathcal{N}(\mu, \Sigma)$ and $Y = BX + \varepsilon$ for a matrix $B \in \mathbf{R}^{p\times m}$ and $\mathbf{R}^p \ni \varepsilon \sim \mathcal{N}(0, \Omega)$ independent of X where $\Omega \in \mathbf{R}^{p\times p}$. Then, as $\mathbb{E}Y = B\mu$, $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)^T = \Sigma B^T$ and $\mathrm{Cov}(Y) = B\Sigma B^T + \Omega$, we have

$$\mathbb{E}(X|Y = y) = \mu + K(y - B\mu)$$

and

$$\mathrm{Cov}(X|Y = y) = \Sigma - K\Sigma K^T = (I - KB)\Sigma,$$

as long as $B\Sigma B^T + \Omega$ is non-singular. Here $K = \Sigma B^T \left(B\Sigma B^T + \Omega\right)^{-1}$.

(iii) If Σ_{XX} is non-singular, then $Y - BX$ is independent of X for $B = \Sigma_{YX}\Sigma_{XX}^{-1}$ and we may write

$$Y = BX + v + \eta$$

for an $\eta \sim \mathcal{N}(0, \Omega)$ with covariance matrix $\Omega = \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$ independent of X, and $v = \mu_Y - B\mu_X$.

(iv) Suppose that (X, Y, Z) is jointly normal mean μ and covariance matrix Σ, partitioned in the same way. If the conditional distribution of X given $Y = y$ and $Z = z$ is given by

$$X|Y = y, Z = z \sim \mathcal{N}(Ky + Gz + \delta, \Xi),$$

then the conditional distribution of X given only $Y = y$ is

$$X|Y = y \sim \mathcal{N}\left(Ky + G\mu_{Z|Y=y} + \delta, \Xi + G\,\mathrm{Cov}(Z|Y)G^T\right).$$

Remark 3.3 (generalized inverse)**.** If Σ_{YY} in Lemma 3.1 (i) is singular, the statement remains true if we choose as Σ_{YY}^{-1} a generalized inverse of Σ_{YY}, see (Rao, 2002, 8.a Note 3). A generalized inverse for a matrix $A \in \mathbf{R}^{m\times p}$ is any matrix $A^- \in \mathbf{R}^{m\times p}$ such that $AA^-A = A$. Given a singular value decomposition $A = UDV^T$, we may obtain the Moore-Penrose inverse $A^\dagger = VD^-U^T$ of A, which is a generalized inverse of A, by inverting the non-zero diagonal elements of D.

Proof. For the first statement, we refer the reader to (Durbin and Koopman, 2012, Chapter 4, Lemma 1).

The second statement follows from substituting the value of K.

The third statement follows from noting that $Y - BX = \begin{pmatrix} -B & I \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}$ follows a Gaussian distribution. A quick calculation reveals that

$$\operatorname{Cov}(Y - BX, X) = \Sigma_{YX} - B\Sigma_{XX} = \Sigma_{YX} - \Sigma_{YX} = 0,$$

showing the independence. Thus $\eta = Y - BX - \delta$ follows a centered Gaussian distribution and equating covariance matrices, we see that Ω has the desired form.

For the final statement, notice that $\xi = X - KY - GZ - \delta$ fulfills

$$\xi | Y = y, Z = z \sim \mathcal{N}(0, \Xi)$$

which does not depend on y or z. Thus the unconditional distribution of ξ is $\mathcal{N}(0, \Xi)$ as well, and ξ is independent of (Y, Z). Rewriting X in terms of Y, Z and ξ, we obtain

$$X = KY + GZ + \delta + \xi,$$

and so

$$\mathbb{E}(X|Y = y) = Ky + G\mathbb{E}(Z|Y = y) + \delta,$$

as well as

$$\begin{aligned}
\operatorname{Cov}(X|Y = y) &= \operatorname{Cov}(KY + GZ + \delta + \xi | Y = y) \\
&= \operatorname{Cov}(GZ + \xi | Y = y) \\
&= \operatorname{Cov}(GZ + \xi) - \operatorname{Cov}(GZ + \xi, Y)\Sigma_{YY}^{-1} \operatorname{Cov}(Y, GZ + \varepsilon) \\
&= G\Sigma_{ZZ}G^T + \Xi - G\Sigma_{ZY}\Sigma_{YY}^{-1}\Sigma_{YZ}G^T \\
&= \Xi + G \operatorname{Cov}(Z|Y)G^T.
\end{aligned}$$

□

After having observed $Y = y$, our main interest lies in the conditional distribution of states X given $Y = y$, which we could obtain by applying Lemma 3.1, i.e. where $B = \text{block-diag}(B_0, \ldots, B_n)$ and $\Omega = \text{block-diag}(\Omega_0, \ldots, \Omega_n)$. However, this would require inversion of the $(n+1)p \times (n+1)p$ matrix $(B\Sigma B + \Omega)$ which becomes numerical infeasible quickly. Instead, we can exploit the sequential structure of the GLSSM, which will allow us to perform conditioning on only a single observation at a time.

To this end, let us denote by $\hat{X}_{t|s}$ the conditional expectation of X_t given a set of observations $y_{:s}$ and by $\Xi_{t|s}$ the conditional covariance matrix of X_t given $Y_{:s} = y_{:s}$. Then

$$X_t | Y_{:s} = y_{:s} \sim \mathcal{N}\left(\hat{X}_{t|s}, \Xi_{t|s}\right).$$

For a given t, three values of s are of particular interest: If $s = t - 1$ determining this conditional distribution is called a **prediction problem**, if $s = t$ this is a **filtering problem** and if $s = n$ a **smoothing problem**, and we call the distributions we seek the **predictive, filtering** or **smoothing distribution** respectively. Similarly we define $\hat{Y}_{t|s} = \mathbb{E}(Y_t | Y_{:s} = y_{:s})$ to be the conditional expectation of Y_t given $Y_{:s} = y_{:s}$, note that $\hat{Y}_{t|s} = Y_t$ if $s \geq t$. Finally, let $\Psi_{t|s} = \operatorname{Cov}(Y_t | Y_{:s} = y_{:s})$ be the conditional covariance matrix of Y_t given $Y_{:s} = y_{:s}$. Again $\Psi_{t|s} = 0$ if $s \geq t$.

These distributions may be obtained efficiently using the celebrated Kalman filter (Algorithm 1) and smoother (Algorithm 2) algorithms, which we state here for completeness.

Algorithm 1 Kalman filter, with runtime $\mathcal{O}(n(m^2+p^3))$

Require: GLSSM (Definition 3.2), observations $y_0, \dots, y_n$.

$A_{-1} \leftarrow I \in \mathbf{R}^{m\times m}$ ▷ Identity Matrix
$u_{-1} \leftarrow \mathbf{0} \in \mathbf{R}^m$
$\hat{X}_{-1|-1} \leftarrow \mathbb{E}X_0$
$\Xi_{0|-1} \leftarrow \mathbf{0}_{m\times m}$
$\ell_{-1} \leftarrow 0$
for $t \leftarrow 0, \dots, n$ **do**
  $\hat{X}_{t|t-1} \leftarrow A_{t-1}\hat{X}_{t-1|t-1} + u_{t-1}$ ▷ prediction
  $\Xi_{t|t-1} \leftarrow A_{t-1}\Xi_{t-1|t-1}A_{t-1}^T + \Sigma_t$
  $\hat{Y}_{t|t-1} \leftarrow B_t\hat{X}_{t|t-1} + v_t$
  $\Psi_{t|t-1} \leftarrow B_t\Xi_{t|t-1}B_t^T + \Omega_t$
  $K_t \leftarrow \Xi_{t|t-1}B_t^T\Psi_{t|t-1}^{-1}$ ▷ filtering
  $\hat{X}_{t|t} \leftarrow \hat{X}_{t|t-1} + K_t(y_t - \hat{Y}_{t|t-1})$
  $\Xi_{t|t} \leftarrow \Xi_{t|t-1} - K_t\Psi_{t|t-1}K_t^T$
  $\ell_t \leftarrow \ell_{t-1} + \frac{p}{2}\log(2\pi) + \frac{1}{2}\log\det\Psi_{t|t-1} + \frac{1}{2}\left(y_t - \hat{Y}_{t|t-1}\right)^T\Psi_{t|t-1}^{-1}\left(y_t - \hat{Y}_{t|t-1}\right)$ ▷ NLL
end for

In Algorithm 1 every time point $t = 0, \dots, n$ is processed in the same way, with a two-step procedure: first we predict the new observation Y_t based on $Y_{:t-1}$. Using the linearity of the system as well as the assumed conditional independence, this is achieved by applying the system dynamics to the current conditional expectation and covariance matrices. After Y_t has been observed, we can update the conditional distribution of the states by appealing to Lemma 3.1. For a rigorous derivation of the Kalman filter, we refer the reader to (Durbin and Koopman, 2012, Chapter 4) or the excellent monograph of (Schneider, 1986).

The Kalman filter is very efficient: each loop iteration requires inversion of the $p \times p$ matrix $\Psi_{t|t-1}$. Assuming this operation dominates the time complexity, e.g. because $m \approx p$, the time complexity of the Kalman filter is $\mathcal{O}(n\,m^3)$, a drastic improvement over the naïve $\mathcal{O}(n^3\,m^3)$, obtained by applying Lemma 3.1 to the joint distribution of (X, Y). Similarly, the space complexity of Algorithm 1 is $\mathcal{O}\left(n\left(m^2+p^2\right)\right)$, and grows only linearly in the number of time steps n.

Notice that the Kalman filter iteratively calculates the negative log-likelihood ℓ_t

$$\ell_t = -\log p(y_{:t}) = -\log\sum_{s=0}^{t}\log p(y_s|y_{:(s-1)})$$

while filtering. This is possible because of the dependency structure of the GLSSM, which makes the increments in ℓ_t tractable, as

$$Y_s|Y_{:(s-1)} \sim \mathcal{N}\left(\hat{Y}_{s|s-1}, \Psi_{s|s-1}\right),$$

for $s = 0, \dots, n$, which is shown in the derivation of the Kalman filter. Thus, the Kalman filter enables us to perform MLE by giving us access to ℓ_n.

historical comment

Depending on the situation at hand, one of the many variants of the basic algorithm presented in Algorithm 1 may be used. If the inversion of $\Psi_{t|t-1}$ is numerically unstable, the filtered covariance matrices $\Xi_{t|t}$ may become numerically non-positive definite. In this case, the square root filter and smoother (Morf and Kailath, 1975) may be used. It is based on Cholesky roots of the involved covariance matrices, ensuring them to be PSD.

When the dimension of observations is much larger than that of the states, $p \gg m$, the information filter (D. Fraser and Potter, 1969) can be used. Instead of performing operations on the covariance matrices, i.e. $\Xi_{t|t-1}$ and $\Psi_{t|t-1}$, the information filter operates on their inverses, the precision

matrices $\Xi_{t|t-1}^{-1}$ and $\Psi_{t|t-1}^{-1}$ as well as rescaled states $\Xi_{t|t-1}^{-1}\hat{X}_{t|t-1}$ and observation $\Psi_{t|t-1}^{-1}\hat{Y}_{t|t-1}$ estimates. This makes the filtering step more efficient, as the computationnaly most intensive step is the calculation of $\Psi_{t|t-1}^{-1}$. However the price one pays is that the prediction step now requires inversion of a $m \times m$ matrix, and as such the computational gains only set in when p is sufficiently large compared to m Assimakis, Adam, and Douladiris, 2012. Note that for the models we consider in Chapter 4 this is usually not the case.

check that this really holds

If the dimensions of the model are so large that calculating the $m \times m$ and $p \times p$ covariance matrices becomes an issue, the simulation based Ensemble Kalman filter (EnKF) (Evensen, 1994) can be used. Instead of calculating the covariance matrices analytically, the EnKF stores a particle approximation to the Gaussian filtering distribution and iteratively performs a prediction and update step with a particle approximation, similar to the analytical update the Kalman filter performs. Despite being based on linear Gaussian dynamics, the EnKF is successfully employed in many high-dimensional non-linear non Gaussian problems (Katzfuss, Stroud, and Wikle, 2016).

For non-linear problems of moderate dimension, i.e. those where we replace the right-hand side of both state (Equation (3.4)) and observation (Equation (3.5)) equations by non-linear functions, other variants such as the Extended Kalman filter (EKF) (Jazwinski, 1970) and the unscented Kalman filter (UKF) (Julier and Uhlmann, 1997) may be used. The EKF applies the Kalman filter to a linearization of the non-linear system around the current conditional means $\hat{X}_{t|t-1}$ and $\hat{X}_{t|t}$. If the systems dynamics are highly non-linear, this approximation can fail. Alternatively, the UKF, which is based on the unscented transform, directly approximates the predicted means and covariance matrix, by constructing a set of deterministic points that are propagated through the systems dynamics.

more on this

In the context of COVID-19, variants of the Kalman filter have been employed to analyse the time-varying behavior of epidemiological parameters. Usually the models start from some theoretical, e.g. compartmental, model of how the epidemic spreads. After time-discretization and possibly linearization, one ends up with a GLSSM, to which the Kalman filter or one of its variants may be applied. In (Arroyo-Marioli et al., 2021) the authors construct a simple GLSSM to reconstruct the time-varying reproduction number from observed growth factors, exploiting the linear relationship between the two quantities in the SIR compartmental model and using the Kalman filter and smoother to perform inference. (Song et al., 2021; Zhu et al., 2021) directly apply the EKF to time-discretized compartmental models, fitting them either to simulated (Zhu et al., 2021) or real (Song et al., 2021) data. Similarly, (Engbert et al., 2020) use the EnKF to fit a stochastic compartmental model to German regional data, where the EnKF allows to deal with the non-linear and non-Gaussian properties on these small spatial scales.

algorithmen konsistent mit gets und =, return value

algorithmen konsistent t, t -1

Algorithm 2 Kalman smoother. Note that the Kalman filter already outputs the smoothed last state $\hat{X}_{n|n}$ and covariance $\Xi_{n|n}$.

Require: GLSSM (Definition 3.2), outputs from Kalman filter (Algorithm 1)

for $t \leftarrow n-1, \dots, 0$ **do**

$G_t = \Xi_{t|t} A_t \Xi_{t+1|t}^{-1}$

$\hat{X}_{t|n} = \hat{X}_{t|t} + G_t \left(\hat{X}_{t+1|n} - \hat{X}_{t+1|t} \right)$

$\Xi_{t|n} = \Xi_{t|t} - G_t \left(\Xi_{t+1|t} - \Xi_{t+1|n} \right) G_t^T$

end for

The Kalman smoother (Algorithm 2) computes the marginal distributions $X_t|Y$ for $t = 0, \dots, n$. Upon closer inspection, the mean and covariance updates resemble that of the Kalman filter

(Algorithm 1). This is no coincidence: By the assumed dependence structure, we obtain the following lemma, which will allow us to prove the recursions.

Lemma 3.2 (conditional indepdence from future observations)**.** *Let $t \in \{0, \dots, n-1\}$ and $s > t$. In a GLSSM, conditional on X_{t+1}, X_t is independent of Y_s, $s > t$.*

Proof. As $s > t$, we have

$$p(x_t, y_s|x_{t+1}) = p(y_s|x_{t+1}, x_t)p(x_t|x_{t+1}) = p(y_s|x_{t+1})p(x_t|x_{t+1})$$

where the second equality follows from the dependency structure of the model. □

We can now sketch the proof for the Kalman smoother recursions, based on the arguments in (Chopin and Papaspiliopoulos, 2020, Chapter 7.3). By the preceding lemma, the conditional distribution of X_t given $Y_{:n}$ and X_{t+1} is the same as that given $Y_{:t}$ and X_{t+1}. We may now regard $X_{t+1} = A_t X_t + u_t + \varepsilon_{t+1}$ as an additional observation at time t, and use the Kalman filter update to determine this conditional distribution:

$$X_t|Y_{:n} = y_{:n}, X_{t+1} = x_{t+1} \sim \mathcal{N}\left(\hat{X}_{t|t} + G_t(x_{t+1} - \hat{X}_{t+1|t}), \Xi_{t|t} - G_t \Xi_{t+1|t} G_t^T\right).$$

As $\hat{X}_{t|t}$ and $\hat{X}_{t+1|t}$ are linear functions of $Y_{:n}$ (actually $Y_{:t}$), we may apply the last statement of Lemma 3.1, to see that, conditional on $Y_{:n} = y_{:n}$, the distribution of X_t is Gaussian with mean

$$\hat{X}_{t|t} + G_t\left(\hat{X}_{t+1|n} - \hat{X}_{t+1|t}\right)$$

and covariance matrix

$$\Xi_{t|t} - G_t \Xi_{t+1|t} G_t^T + G_t \Xi_{t+1|n} G_t^T = \Xi_{t|t} - G_t\left(\Xi_{t+1|t} - \Xi_{t+1|n}\right) G_t^T.$$

These quantities are calculated by the Kalman smoother (Algorithm 2).

Going back to the proof of the last statement in Lemma 3.1, we see that we may actually write

$$X_t = \hat{X}_{t|t} + G_t(X_{t+1} - \hat{X}_{t+1|t}) + \xi_t, \tag{3.6}$$

for a $\xi_t \sim \mathcal{N}\left(0, \Xi_{t|t} - G_t \Xi_{t+1|t} G_t\right)$ which is independent of $Y_{:n}$ and X_{t+1}. This recurrence may be used to generate samples from the joint smoothing distribution, which is useful if one is interested in non-linear functionals of the smoothing distribution that involve multiple states at once, such as a moving median or maximum. It is based on the following decomposition of the smoothing density

$$p(x|y) = p(x_n|y) \prod_{t=n-1}^{0} p(x_t|x_{t+1}, y_{:t}).$$

The resulting algorithm is called the Forwards Filter, Backwards Sampling (FFBS) (Algorithm 3) and was first described in (Frühwirth-Schnatter, 1994) in the context of a data augmentation algorithm for Bayesian analysis of GLSSM. In this paper, the hyperparameters θ follow an inverse gamma distribution, and one is interested in the posterior marginals of $p(\theta|Y)$, e.g. to determine the posterior marginals of states $p(X|Y)$ using MC-integration.

Remark 3.4 (regularity of Σ_t and Ω_t)**.** Throughout this section, we have assumed, either explicitly or implicitly, that the innovation and observation covariance matrices Σ_t and Ω_t are non-singular, i.e. SPD.

For the Kalman filter we require that for every t, $\Psi_{t|t-1}$ is non-singular, i.e. that we can apply Lemma 3.1 (i). This is fulfilled as soon as Ω_t is non-singular, which is a reasonable assumption in most models. Following the remark after Lemma 3.1, we could also replace $\Psi_{t|t-1}^{-1}$ in Algorithm 1 by its Moore-Penrose inverse.

A similar argument can be made for singular $\Xi_{t+1|t}$, where we replace $\Xi_{t+1|t}^{-1}$ by its Moore-Penrose inverse in the Kalman smoother (Algorithm 2) and the FFBS (Algorithm 3).

Algorithm 3 Forwards filter, backwards smoother (Frühwirth-Schnatter, 1994, Proposition 1)

Require: GLSSM (Definition 3.2), outputs from Kalman filter (Algorithm 1)

1: Simulate $\check{X}_{n|n} \sim \mathcal{N}(\hat{X}_{n|n}, \Xi_{n|n})$
2: **for** $t \leftarrow n-1, \ldots, 0$ **do**
3: $\quad G_t = \Xi_{t|t} A_t \Xi_{t+1|t}^{-1}$
4: $\quad$ Simulate $\xi_t \sim \mathcal{N}(0, \Xi_{t|t} - G_t \Xi_{t+1|t} G_t^T)$
5: $\quad$ Set $\check{X}_{t|n} = \hat{X}_{t|t} + G_t \left(\check{X}_{t+1} - \hat{X}_{t+1|t} \right) + \xi_t$
6: **end for**

The attractive feature of GLSSMs is that a large part of inference is analytically feasible: we may calculate the likelihood, smoothing distribution and sample from it. However, the modeling capacity of GLSSMs is limited: most interesting phenomena in the context of this thesis follow neither linear dynamics nor are well modeled by a Gaussian distribution, see also the discussion in Section 3.1.

Nevertheless, linearization of non-linear dynamics suggests that GLSSMs may have some use as approximations to these more complicated phenomena, provided they are sufficiently close to Gaussian models, e.g. unimodal and without heavy tails. We start to move away from linear Gaussian models by allowing observations that are non-Gaussian.

3.3 Partially Gaussian state space models

The distribution of observations is never Gaussian - all we may hope for is that the data-generating mechanism is close enough to a Gaussian distribution that inferences made in an GLSSM may carry over. For epidemiological models, Gaussian distributions may be appropriate if incidences are high, e.g. during large outbreaks in a whole country. When case numbers are small, the discrete nature of incidences is better captured by a distribution on $\mathbf{N}_0$, and standard distributions used are the Poisson and negative binomial distributions, see e.g. (Lloyd-Smith et al., 2005), see also the discussion in Section 3.1. We thus want SSMs where observations are allowed to follow these non-Gaussian distributions.

Concerning the distribution of states, we keep the linear Gaussian assumption, i.e. Equation (3.4). As argued in Section 3.1,

do this there

using Gaussian states and transitions allows for flexible modeling of many epidemiological desiderata. Furthermore, keeping the states Gaussian will enable us to use Efficient Importance Sampling (EIS) effectively, by constructing approximations via GLSSM which possess the same state dynamics. Alternatively, t-distributed innovations or more general transition kernels could be employed and we refer the interested reader to (Durbin and Koopman, 2012, Part II) for a selection of these models. The following definition is that of (Koopman, Lit, and Nguyen, 2019), which itself is an extension of earlier work of (Shephard, 1994). (Shephard, 1994) considered only SSMs where, conditional on another Markov process $Z = (Z_t)_{t=0,\ldots,n}$, model is a full GLSSM. While this formulation allows for efficient inference if the distribution of Z leads to a tractable conditional distribution $Z|(X,Y)$. As their definition involves a conditional GLSSM, the observations still take values in $\mathbf{R}^p$, not $\mathbf{N}^p$ as is necessary for our endeavors. Thus we opt for the definition presented in (Koopman, Lit, and Nguyen, 2019), where we replace the Gaussian observations (Equation (3.5)) with arbitrary distributions.

Definition 3.3 (Partially Gaussian state space model (PGSSM))**.** A Partially Gaussian state space model (PGSSM) is a joint distribution for (X, Y) where states X follow Equation (3.4), i.e.

$$X_{t+1} = A_t X_t + u_t + \varepsilon_{t+1} \qquad t = 0, \ldots, n-1,$$

with $X_0 \sim \mathcal{N}(0, \Sigma_0)$, $\varepsilon_t \sim \mathcal{N}(0, \Sigma_t)$ for $t = 1, \ldots, n$ and X_0, $(\varepsilon_t)_{t=1,\ldots,n}$ jointly independent.

Furthermore, the observations Y form a conditional Markov process, conditional on states X, of the following form:

$$p(y|x) = \prod_{t=0}^{n} p(y_t|x_t, y_{t-1}).$$

Here $p(y_t|x_t, y_{t-1})$ are allowed to take any arbitrary distribution[1].

It is straightforward to check that a PGSSM is indeed a SSM.

Remark 3.5. Recalling Remark 3.1, if our main interest lies in the conditional distribution $X|Y = y$ for a fixed set of observations y, it will suffice to consider models where

$$p(y|x) = \prod_{t=0}^{n} p(y_t|x_t)$$

holds, and we will do so in the following to enhance readability. At points where this distinction matters, e.g.

add example

, we will give appropriate remarks.

Both the Poisson and negative binomial belong to the class of exponential family distributions. As such, their densities have a convenient structure, allowing only for a linear interaction between the natural parameter and the densities argument. We refer to (Brown, 1986) for a comprehensive treatment of exponential families and use their definitions throughout this section.

Definition 3.4 (exponential family)**.** Let μ be a σ-finite measure on $\mathbf{R}^p$ and denote by

$$\Psi = \left\{ \psi \in \mathbf{R}^p : \int \exp\left(\psi^T y\right) \, \mathrm{d}\mu(y) < \infty \right\}$$

the set of parameters ψ such that the moment-generating function of μ is finite. For every $\psi \in \Psi$

$$p_\psi(y) = Z(\psi)^{-1} \exp(\psi^T y)$$

defines a probability density with respect to the measure μ, where

$$Z(\psi) = \int \exp\left(\psi^T x\right) \, \mathrm{d}\mu(y)$$

is the normalizing constant. We call both the densities p_ψ and induced probability measures

$$\mathbf{P}_\psi(A) = \int_A p_\psi(y) \, \mathrm{d}\mu(y),$$

for measurable $A \subset \mathbf{R}^p$, a **standard exponential family**.

Conversely, let $\mathbf{P}_\psi, \psi \in \Psi$ be a given parametric family of probability measures on some space $\mathcal{Y}$ that is absolutely continuous with respect to a common dominating measure μ. Suppose there exist a reparametrization $\eta : \Psi \to \mathbf{R}^p$, a statistic $T : \mathcal{Y} \to \mathbf{R}^p$ and functions $Z : \Psi \to \mathbf{R}$, $h : \mathcal{Y} \to \mathbf{R}$, such that

$$p_\psi(y) = \frac{\mathrm{d}\mathbf{P}_\psi}{\mathrm{d}\mu} = Z(\psi)h(y) \exp\left(\eta(\psi)^T T(y)\right),$$

then we call $\mathbf{P}_\psi, \psi \in \Psi$ and $p_\psi, \psi \in \Psi$ a **p-dimensional exponential family** and $(\mathbf{P}_\psi)_{\psi \in \Psi}$ a **p-dimensional natural exponential family**. If $\eta(\psi) = \psi$ we call ψ the natural parameter. If $T(y) = y$, we call y the natural observation. By reparametrization (in ψ) and sufficiency (in y) every p-dimensional exponential family can be written as an equivalent standard exponential family, see the elaborations in (Brown, 1986, Chapter 1).

[1] Recall that we have not specified $\mu_\mathcal{Y}$, so it is always possible to use $p = \mathbf{1}_\mathcal{Y}$, the constant function.

Exponential families have the attractive property that they are log-concave in their parameters. As such the Fisher-information is always positive semidefinite, which will be crucial in defining surrogate Gaussian models in Section 3.6.

Lemma 3.3 (log-concavity of exponential family distributions)**.** *Let $p_\psi, \psi \in \Psi$ be a natural p-dimensional exponential family and Ψ convex and open in $\mathbf{R}^p$. In this case $\psi \mapsto \log p_\psi(y)$ is concave for every $y \in \mathbf{R}^p$.*

Proof. As $\log p_\psi(y) = -\log Z(\psi) + \psi^T y$ it suffices to show that $\psi \mapsto \log Z(\psi)$ is convex. However,

$$\psi \mapsto \log Z(\psi) = \log \int \exp\left(\psi^T y\right) \mathrm{d}\mu(y)$$

is the cumulant generating function of the base measure μ, which is convex (Billingsley, 1995, p. 144f). □

Additionally, the moment generating function $\psi \mapsto Z(\psi)$ is smooth on the interior of Ψ and allows to switch the order of integration and differentiation.

Theorem 3.1 ((Brown, 1986, Theorem 2.2, Corollary 2.3))**.** *Let $\psi \in \operatorname{int} \Psi$ be an interior point. Then the moment generating function $Z : \Psi \to \mathbf{R}$ is infinitely often differentiable with derivatives*

$$\frac{\partial^{|\alpha|}}{\partial^\alpha \psi} Z(\psi) = \int y^\alpha \exp\left(\psi^T y\right) \mathrm{d}\mu(y)$$

for any multi-index $\alpha \in \mathbf{N}^k$.

Additionally, the gradient of $\log Z$, $\nabla_\psi \log Z(\psi)$ is given by

$$\nabla_\psi \log Z(\psi) = \mathbb{E}T(X),$$

and the Hessian of $\log Z$, $H_\psi \log Z(\psi)$ by

$$H_\psi \log Z(\psi) = \operatorname{Cov}(T(X)),$$

where $X \sim \mathcal{P}_\psi$.

Example 3.1 (Poisson & negative binomial distribution)**.** Both the family of Poisson distributions, parameterized by rate λ and the negative binomial distribution, parameterized by success probability p with fixed overdispersion r form an exponential family.

The log-density of the Poisson distribution with rate λ, $\mathrm{Pois}(\lambda)$ w.r.t. the counting measure on $\mathbf{N}_0$ is

$$\log p_\lambda(x) = -\lambda + x \log \lambda - \log x!.$$

Thus the Poisson distribution forms an exponential family with natural parameter $\log \lambda$, natural statistic id (the identity), base measure $h(x) = \frac{1}{x!}$ and log-partition function $Z(\lambda) = \exp(-\lambda)$.

The log-density of the negative binomial distribution with overdispersion parameter r and success probability p $\mathrm{NegBinom}(p, r)$ is

$$\log p_{r,p}(x) = \log \binom{x + r - 1}{x} + x \log(1 - p) + r \log p.$$

For fixed r these distributions form an exponential family with natural parameter $\log(1-p)$, natural statistic $T = \mathrm{id}$, base measure $h(x) = \log \binom{x+r-1}{x}$ and log-partition function $Z(p) = r \log p$.

In this parametrization the mean of the $\mathrm{NegBinom}(p, r)$ distribution is $\mu = r\frac{1-p}{p}$ and its variance is $r\frac{1-p}{p^2}$. An alternative parametrization that will become useful Chapter 4 is that by the log mean $\xi = \log \mu$ and overdisperision r. As $p = \frac{r}{r+\mu}$, this parametrization has log-density

$$\log p_{r,\xi}(x) = \log \binom{x + r - 1}{x} + x\xi - (r + x) \log(\exp \xi + r) - r \log r,$$

which does not form a natural exponential family. However, it retains the log-concavity of Lemma 3.3, as a quick calculation reveals that

$$\partial^2_{\xi^2} \log p_{r,\xi}(x) = -(r+x)\frac{r\exp(-\xi)}{(r\exp(-\xi)+1)^2} < 0$$

for all $x \in \mathbf{N}_0$.

The models we study in Chapter 4 belong, for the most part,

check

to the following subclass of PGSSM models.

Definition 3.5 (Exponential Family Partially Gaussian state space model (EGSSM))**.** An Exponential Family Partially Gaussian state space model (EGSSM) is a PGSSM where the conditional distribution of Y_t given X_t comes from an exponential family with respect to a base measure μ_t, i.e.

$$p(y_t|x_t) = h_t(y_t)Z_t(x_t)\exp\left(\eta_t(x_t)^T T_t(y_t)\right)$$

for suitable functions h_t, Z_t, η_t, T_t. If Y_t in the PGSSM is allowed to depend on the previous Y_{t-1}, the functions h_t, Z_t, η_t and T_t may depend on y_{t-1}.

If, additionally, matrices $B_t \in \mathbf{R}^{p\times m}$ exist, such that for the signal $S_t = B_t X_t \in \mathbf{R}^p$, Y_t only depends on X_t through S_t, i.e. it holds

$$p(y_t|x_t) = \prod_{i=1}^{p} h_t^i(y_t^i)\, Z_t^i(s_t)\, \exp\left(\eta_t^i(s_t^i)\, T(y_t^i)\right),$$

for functions $h_t^i : \mathbf{R} \to \mathbf{R}, Z_t^i : \mathbf{R} \to \mathbf{R}, \eta_t^i : \mathbf{R} \to \mathbf{R}, T : \mathbf{R} \to \mathbf{R}$, $i = 1, \dots p$, we say the Logconcave state space model (LCSSM) has a **linear signal**, similar to the treatment in (Durbin and Koopman, 2012, Part II).

Remark 3.6. To simplify notation we will usually assume that the functions h, Z and T are the same for all t (and i, if the LCSSM has a linear signal) and drop in our notation the dependence of h, Z, and T on t (and i). Similarly, we assume that the base measure μ_t is the same for all t.

From Lemma 3.3, we immediately obtain the following results (Durbin and Koopman, 2012, Section 10.6.4)

Lemma 3.4 (log-concavity of the smoothing distribution)**.** *Consider an EGSSM, where $\eta_t = \mathrm{id}$ for all t. Then $x \mapsto \log p(x|y)$ is concave for every a.e. $Y = y$.*

Proof. We may write

$$\log p(x|y) = \log p(y|x) + \log p(x) - \log p(y),$$

where the last term does not depend on x. $\log p(x)$ is concave, as $p(x)$ is the joint density of a multivariate Gaussian distribution. Furthermore

$$\log p(y|x) = \sum_{t=0}^{n} \log p(y_t|x_t, y_{t-1}),$$

which, by Lemma 3.3 is concave in x. □

Notice that the dependence of Y_t on Y_{t-1} does not influence the statement of this lemma, as we are interested in properties of $x \mapsto p(x|y)$.

As in the previous chapter, after having observed Y, one is interested in the conditional distribution of states X, given Y. If the observations are not Gaussian, this is a difficult task as the distribution is not analytically tractable. Instead, approximations, e.g. the Laplace approximation (LA), which will exploit the log-concavity developed here or simulation-based inference, e.g. importance sampling

(Sections 3.4 and 3.6), sequential Monte Carlo (Chopin and Papaspiliopoulos, 2020) or MCMC-methods (Brooks et al., 2011) are used. Similarly, fitting hyperparameters ψ by maximum likelihood inference becomes more difficult as evaluating $\ell(\psi) = p(y) = \int p(x, y)\,\mathrm{d}x$ is not analytically available, thus requiring numerical or simulation methods for evaluation and gradient descent or EM-techniques for optimization, see Section 3.7.

In this thesis, we will focus on importance sampling methods, which are the focus of the next section.

3.4 Importance Sampling

Importance sampling is a simulation technique that allows us to approximate integrals w.r.t a measure of interest, the target, by sampling from a tractable approximation, the proposal, instead, thus performing Monte-Carlo integration. To account for the fact that we did not sample from the correct probability measure, we weight samples according to their importance. As the user has freedom in the choice of approximation (except for some technical conditions), importance sampling also acts as a variance reduction technique with better approximations resulting in smaller Monte-Carlo variance. Thus the role that importance sampling plays is twofold: first, it enables Monte-Carlo integration even if sampling from the target is not possible, and second it allows us to do so in an efficient way by choosing, to be defined precisely below, the approximation in an optimal way.

Alternative approaches to importance sampling for performing inference in SSMs include Markov chain Monte Carlo (MCMC) and SMC. Recall from the introduction to this chapter that this inference concerns three objectives: maximum likelihood estimation, i.e. evaluation and optimization of the likelihood, access to the posterior distribution $X_{:n}|Y_{:n}$ and prediction of future states and observations. Let us give a concise comparison of these alternative approaches, weighing their advantages and disadvantages over importance sampling, in particular for the SSMs that this thesis deals with.

MCMC (Brooks et al., 2011) is a simulation technique that allows to simulation of correlated samples from a target distribution by constructing a Markov chain that has as its invariant distribution the desired distribution. For Metropolis-Hastings MCMC, one needs access to the density of the sought distribution up to a constant to simulate a step in the Markov chain. While this method is very general, it fails in high dimensions and current research in MCMC methods investigates this

quotes

curse of dimensionality

citep something

.

MCMC vs. IS

SMC (Chopin and Papaspiliopoulos, 2020) or particle filters, use sequential importance sampling to provide a particle approximation to the filtering distributions $X_t|Y_{:t}$, essentially decomposing the problem into a n importance sampling steps. To avoid particle collapse, SMC is usually equipped with a resampling step once the effective sample size of the current set of particles drops below a specified level. Once the final filtering distribution $X_n|Y_{:n}$ is approximated, the smoothing distribution may be obtained in several ways ...

look up Chopin

.

Conveniently, SMC allows us to approximate the likelihood $\ell(\theta)$ for a single parameter by a single pass of the particle filter. However, the discrete nature of resampling makes the approximated likelihood non-continuous, complicating maximum likelihood inference. (Chopin and Papaspiliopoulos, 2020, Chapter 14) discusses several strategies: the first amounts to importance sampling of the order as

discussed in this thesis, where one fixes a reference parameter θ_0 to perform importance sampling with $p_{\theta_0}(x|y)$ against $p_\theta(x|y)$. The second strategy only works in the univariate case and consists of approximating the non-continuous inverse CDFs appearing in the resampling step by continuous ones. Finally, if the dependence on the hyperparameters θ allows for application of the EM-algorithm, it may be used to perform the optimization. Contrary to SMC, the global importance sampling approach we discuss in Sections 3.6 and 3.7 allows us to perform

...

This chapter proceeds with a general treatment of importance sampling, loosely based on (Chopin and Papaspiliopoulos, 2020, Chapter 8) and (Durbin and Koopman, 2012, Chapter 11). Subsequently, we will focus our attention on methods to obtain good importance sampling proposals.

Suppose we have a function $h : \mathcal{X} \to \mathbf{R}$ whose integral w.r.t. to some measure μ,

$$\zeta = \int_{\mathcal{X}} h(x)\,\mathrm{d}\mu(x),$$

exists and whose value we want to compute. Furthermore, suppose that we can write

$$\int_{\mathcal{X}} h(x)\,\mathrm{d}\mu(x) = \int_{\mathcal{X}} f(x)\,\mathrm{d}\mathbf{P}(x) = \mathbf{P}[f],$$

operatorschreibeweise everywhere

for a probability measure $\mathbf{P}$ and function $f : \mathcal{X} \to \mathbf{R}$, e.g. because $\mathbf{P} = p\mu$ and $h(x) = f(x)p(x)$ μ-a.s. . Let $\mathbf{G}$ be a another probability measure on $\mathcal{X}$ such that $f\mathbf{P}$ is absolutely continuous with respect to $\mathbf{G}$, $f\mathbf{P} \ll \mathbf{G}$, and let $v = \frac{\mathrm{d}f\mathbf{P}}{\mathrm{d}\mathbf{G}}$ be the corresponding Radon-Nikodym derivative. Then

$$\zeta = \mathbf{P}[f] = \int_{\mathcal{X}} f(x)\,\mathrm{d}\mathbf{P}(x) = \int_{\mathcal{X}} \left(\frac{\mathrm{d}f\mathbf{P}}{\mathrm{d}\mathbf{G}}\right)\,\mathrm{d}\mathbf{G}(x) = \mathbf{G}[v]$$

which suggests to estimate ζ by Monte-Carlo integration:

$$\hat{\zeta} = \frac{1}{N}\sum_{i=1}^{N} v(X^i),$$

the importance sampling estimate of ζ. The importance samples $X^i, i = 1, \ldots, N$ have distribution $\mathbf{G}$, and will usually be i.i.d. For this procedure to work, we want $\hat{\zeta}$ to fulfill a law of large numbers and a central limit theorem, so we will want $v \in L^2(\mathbf{G})$, where $L^p(\nu)$ is the space of p-times ν-integrable functions for a measure ν. The i.i.d. assumption could also be dropped, e.g. when we employ antithetic variables, see (Ripley, 2009, Section 5.3). Here we call $\hat{\zeta}$ the importance sampling estimate of ζ.

If $v \in L^2(\mathbf{G})$ and under i.i.d. sampling the Monte-Carlo variance of $\hat{\zeta}$ is $\frac{\mathrm{Var}\left(v(X^i)\right)}{N}$, and so naturally we want $\mathrm{Var}\left(v(X^i)\right)$ to be small to ensure fast convergence of $\hat{\zeta}$. As v depends on the proposal $\mathbf{G}$, and we have the flexibility to choose $\mathbf{G}$, importance sampling acts as a variance reduction technique.

A classical result is that the minimum MSE proposal $\mathbf{G}^*$ has a closed form. Indeed it is given by the total variation measure of $f\mathbf{P}$, renormalized to be a probability measure, which can be shown by a simple application of Jensen's inequality.

Proposition 3.1 (Chopin and Papaspiliopoulos, 2020, Proposition 8.2)**.** *[minimum MSE proposal] The proposal $\mathbf{G}^*$ that minimizes the MSE of importance sampling is given by*

$$\mathbf{G}^* = \frac{|f|}{\mathbf{P}\left[|f|\right]}\mathbf{P}.$$

Unfortunately, this optimality result has no practical use, indeed if f is positive we would need to obtain $\mathbf{P}[f]$ first, the overall target of our endeavor. Additionally, sampling from $\mathbf{G}^*$ is not guaranteed to be practically feasible.

If the Radon-Nikodym derivative $w = \frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{G}}$ exists, then $v = fw$, which, for the problems we will study, is usually the case. In this case

$$\hat{\zeta} = \frac{1}{N}\sum_{i=1}^{N} f(X^i)w(X^i),$$

where $w(X^i)$ is called the importance weight, or just weight, of the i-th sample. If the samples is clear from the context we sometimes write $w^i = w(X^i)$. This motivates us to regard

$$\hat{\mathbf{P}}_N = \frac{1}{N}\sum_{i=1}^{N} w(X_i)\delta_{X_i}, \tag{3.7}$$

as a particle approximation of $\mathbf{P}$, in the sense that for sufficiently well behaved test functions f, as $N \to \infty$

$$\hat{\mathbf{P}}_N[f] = \frac{1}{N}\sum_{i=1}^{N} f(X^i)w(X^i) \to \mathbf{P}[f].$$

We will return to the question of which functions f to consider further below and assume in the following discussion $fw \in L^2(\mathbf{G})$.

To perform importance sampling one must be able to evaluate w. In the context of PGSSMs this is usually not possible: if $\mathbf{P}$ is the intractable conditional distribution of $X|Y$, then the integration constant of its density $p(y)$ is not analytically available. Still, we can usually evaluate the weights up to a constant, i.e.

$$\tilde{w}(x) \propto \frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{G}}(x)$$

is available. The missing constant is then $\mathbf{G}\tilde{w}$, which is itself amenable to importance sampling: we may estimate it by $\sum_{i=1}^{N}\tilde{w}(X^i)$. This leads to the so-called self-normalized importance sampling weights

$$W_i = \frac{w(X^i)}{\sum_{i=1}^{N} w(X^i)},$$

Monte Carlo estimates

$$\hat{\zeta} = \sum_{i=1}^{N} W_i f(X^i),$$

and particle approximation

$$\hat{\mathbf{P}}_N = \sum_{i=1}^{N} W_i \delta_{X^i}.$$

Unless $\tilde{w}$ is degenerate, i.e. constant,

$$\hat{\zeta} = \frac{\sum_{i=1}^{N}\tilde{w}(X^i f(X^i))}{\sum_{i=1}^{N}\tilde{w}(X^i)}$$

is a ratio of two non-constant, unbiased estimators and so is itself biased. Nevertheless, noticing that the rescaled denominator $\frac{1}{N}\sum_{i=1}^{N}\tilde{w}(X^i)$ consistently estimates the integration constant $\mathbf{G}\tilde{w}$, allows us to apply Slutsky's lemma and obtain a central limit theorem for $\hat{\zeta}$ (recall that we assumed $fw \in L^2(\mathbf{G})$).

The class for test functions f for which this holds depends on $\mathbf{P}$ and $\mathbf{G}$. (Agapiou et al., 2017) study the behavior of uniformly bounded test functions $\|f\| \leq 1$. For these functions it suffices that $w \in L^2(\mathbf{G})$ to ensure asymptotic normality of ζ. Thus an important quantity is

$$\rho = \frac{1}{(\mathbf{G}\tilde{w})^2}\mathbf{G}[\tilde{w}^2] = \mathbf{G}[w^2] = \mathbf{P}[w],$$

the second moment of the importance sampling weights. (Agapiou et al., 2017) show that the bias

$$\left|\mathbb{E}(\hat{\mathbf{P}}_N - \mathbf{P})[f]\right|$$

and mean-squared error (MSE)

$$\mathbb{E}\left((\hat{\mathbf{P}}_N - \mathbf{P})[f]\right)^2$$

of importance sampling are both, for bounded f, of order $\mathcal{O}\left(\frac{\rho}{N}\right)$. Here the expectation $\mathbb{E}$ is with respect to the random particles $X^1, \ldots, X^N$. Consequently, for bounded functions, keeping $\frac{\rho}{N}$ small produces importance sampling estimates with small bias and MSE. This can be achieved in two ways: either we choose $\mathbf{G}$ „close enough“ to $\mathbf{P}$ to ensure small ρ, or we choose N large enough to compensate for a large ρ.

Applying Jensen's inequality, we see that

$$\mathcal{D}_{\mathrm{KL}}(\mathbf{P}||\mathbf{G}) = \mathbf{P}[\log w] \leq \log \mathbf{P}[w] = \log \rho,$$

so small ρ implies a small KL-divergence as well. Conversely, the following theorem of Chatterjee and Diaconis implies that a small KL-divergence is both sufficient and necessary for importance sampling to perform well.

Theorem 3.2 (Chatterjee and Diaconis, 2018, Theorem 1.1)**.** *Let $\mathbf{P}$ and $\mathbf{G}$ be probability measures on a measurable space $(\mathcal{X}, \mathcal{B})$ such that $\mathbf{P} \ll \mathbf{G}$ and let $f \in \mathbf{L}^2(\mathbf{P})$ be a function with $\|f\|_{L^2(\mathbf{P})} = \left(\mathbf{P}f^2\right)^{1/2} < \infty$. Let Y be an $\mathcal{X}$ valued random variable with law $\mathbf{P}$.*

Let $L = \mathcal{D}_{KL}(\mathbf{P}||\mathbf{G}) = \mathbb{E}\log w(Y)$ be the KL-divergence between $\mathbf{P}$ and $\mathbf{G}$, and let

$$\hat{\mathbf{P}}_N = \sum_{i=1}^{N} w(X^i)\delta_{X^i}$$

be the particle approximations of $\mathbf{P}$ based on samples $X^1, \ldots, X^N \overset{i.i.d}{\sim} \mathbf{G}$, $N \in \mathbf{N}$.

If the sample size N is given by $N = \exp(L + t)$ for a $t \geq 0$,

$$\mathbb{E}\left|\hat{\mathbf{P}}_N[f] - \mathbf{P}[f]\right| \leq \|f\|_{L^2(\mathbf{P})}\left(\exp(-t/4) + 2\sqrt{\mathbb{P}\left(\log w(Z) > L + t/2\right)}\right). \tag{3.8}$$

Conversely, if $N = \exp(L - s)$ for $s \geq 0$, then for any $\delta \in (0, 1)$

$$\mathbb{P}(\hat{\mathbf{P}}_N[\mathbf{1}] \geq 1 - \delta) \leq \exp\left(-\frac{s}{2}\right) + \frac{\mathbb{P}\left(\log w(Z) \leq L - \frac{s}{2}\right)}{1 - \delta}, \tag{3.9}$$

where $\mathbf{1}$ is the constant function $x \mapsto 1$.

Notice the boldface $\mathbb{P}$ and $\mathbb{E}$ to differentiate the measures $\mathbf{P}$ and $\mathbf{G}$ from expectations and probabilities with respect to the abstract probability space $(\Omega, \mathcal{A}, \mathbb{P})$ where the random variables $X_1, \ldots, X_N$ and Y live.

The proof of this theorem is based on splitting $\mathcal{X}$ into $\{\log w \leq L + \frac{t}{2}\}$ and its complement and straightforward, it may be found in the Appendix of (Chatterjee and Diaconis, 2018). Theorem 1.2 in the same paper provides a qualitatively similar result for autonormalised importance sampling.

Let us consider the implications of Theorem 3.2, starting with Equation (3.8), by devising heuristics to decide when $\mathbf{G}$ is a good proposal for fixed sample size N, and assume for simplicity that $\|f\|_{L^2(\mathbf{P})} = 1$. First of all, as $t = \log N - L$, we have $\exp(-t/4) = \frac{\exp(L/4)}{N^{\frac{1}{4}}}$, so for large N this term becomes negligible, and the interesting term in inequality (3.8) is the second one. As $\mathbb{E}\log w(Z) = L$, this term is a tail probability and we can use standard mass-concentration inequalities to analyze its behavior as t (and so N) grows. Markov's inequality tells us that

$$\mathbb{P}\left(\log w(Z) > L + \frac{t}{2}\right) \leq \frac{L}{L + t/2} = \frac{2}{1 + \frac{\log N}{L}}.$$

Second, if, additionally, $\log w(Z)$ has finite variance, Chebyshev's inequality yields

$$\mathbb{P}\left(\log w(Z) > L + \frac{t}{2}\right) \leq \frac{4\operatorname{Var}(\log w(Z))}{t^2} = \frac{4\operatorname{Var}(\log w(Z))}{(\log N - L)^2}.$$

In both upper bounds provided by the concentration inequalities, all else being equal, a smaller KL-divergence will yield a tighter bound. However, in Chebyshev's inequality, the variance of log weights also plays a role, and will surely be different for different proposals. Assuming $\mathbf{G} \ll \mathbf{P}$, we have $\frac{\mathrm{d}\mathbf{G}}{\mathrm{d}\mathbf{P}} = \frac{1}{w}$ and so

$$\mathbb{E}\exp(-\log w(Z)) = \mathbb{E}\frac{1}{w(Z)} = \mathbf{P}\left[\frac{\mathrm{d}\mathbf{G}}{\mathrm{d}\mathbf{P}}\right] = 1,$$

If the log-weights are bounded from above and below, the following lemma shows that as the variance of $U = -\log w(Z)$ goes to 0, their mean,

$$\mathbb{E}U = \mathbb{E} - \log w(Z) = -\mathcal{D}_{\mathrm{KL}}(\mathbf{P}||\mathbf{G})$$

goes to 0 as well.

Lemma 3.5. *For $a, b \in \mathbf{R}$, let $U \in [a, b]$ be a bounded random variable with variance σ^2 and $\mathbb{E}\exp U = 1$. Let $\mu = \mathbb{E}U$ be the mean of U. Then there exists a $\delta \in [a, b]$, such that*

$$0 \geq \mu = \log\left(1 - \delta\frac{\sigma^2}{2}\right).$$

If, additionally, $\sigma^2 < \frac{2}{b}$ then

$$\mu \geq \log\left(1 - b\frac{\sigma^2}{2}\right).$$

Proof. As U is bounded, all involved expectations exist and are finite. That $\mu \leq 0$ follows from Jensen's inequality. We perform a first-order Taylor expansion of $\exp(U - \mu)$, where the random variable ξ is between $U - \mu$ and 0:

$$1 = \exp(\mu)\mathbb{E}\exp(U - \mu) = \exp(\mu)\left(1 + \mathbb{E}(U - \mu) + \mathbb{E}\left(\frac{(U - \mu)^2}{2}\exp(\xi)\right)\right).$$

Then $\xi' = \xi + \mu \in [a, b]$, and note that, unless $U = 1$ a.s., $\mathbb{E}\exp = 1$ forces $a < 0 < b$. Thus

$$1 = \exp(\mu) + \mathbb{E}\left(\frac{(U - \mu)^2}{2}\exp(\xi')\right),$$

and as $\xi' \in [a, b]$, the expectation is in $\left[\exp(a)\frac{\sigma^2}{2}, \exp(b)\frac{\sigma^2}{2}\right]$, i.e. $\mathbb{E}\left(\frac{(U-\mu)^2}{2}\exp(\xi')\right) = \delta\frac{\sigma^2}{2}$ for some $\delta \in [a, b]$. Solving for μ, we get

$$\mu = \log\left(1 - \delta\frac{\sigma^2}{2}\right),$$

as promised.

The second statement follows from $\delta \leq b$ and the monotonicity of log, where the condition ensures that the argument is positive. □

Corollary 3.1. *Let $\mathbf{P}$ and $\mathbf{G}$ be equivalent probability measures with bounded Radon-Nikodym derivative $w = \frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{G}} \in [a, b]$, $a, b \in \mathbf{R}$ and KL-divergence $\mathcal{D}_{KL}(\mathbf{P}||\mathbf{G}) = \mathbf{P}[\log w]$.*

If $\log w \in L^2(\mathbf{P})$ with variance $\sigma^2 = \mathbf{P}[(\log w - L)^2]$, and $\sigma^2 < \frac{2}{b}$, then

$$\mathcal{D}_{KL}(\mathbf{P}||\mathbf{G}) \leq -\log\left(1 - \exp(b)\frac{\sigma^2}{2}\right).$$

Under the assumptions of this corollary, we see that a small variance of the log-weights implies a small KL-divergence, which in turn implies good importance sampling performance.

Let us now discuss the implications of Equation (3.9). We see that for large s, i.e. $N \ll \exp(L)$, the right-hand side is small, and so the probability that importance sampling fails for the constant function is practically relevant. Observe that here

$$\hat{\mathbf{P}}_N[\mathbf{1}] = \frac{1}{N}\sum_{i=1}^{N} w_i$$

is the mean of weights, which, for the standard weights w, does not have to sum to 1. As a result, Chatterjee and Diaconis recommend to choose $N = \mathcal{O}(\exp(\mathcal{D}_{\mathrm{KL}}(\mathbf{P}||\mathbf{G})))$.

Based on this discussion, we see that choosing $\mathbf{G}$ such that either the KL-divergence or the variance of the log-weights is small is sensible. Making the variance small has the additional advantage that it, at least for bounded log-weights, also implies an upper bound for the KL-divergence. We will return to this train of thought when we discuss optimal ways of performing importance sampling, such as the CE-method (minimizing the KL-divergence) and EIS (minimizing the variance of log-weights) in the following sub-chapters.

In practice, we will want to judge whether for an actual sample $X^1, \dots X^N \overset{\text{i.i.d}}{\sim} \mathbf{G}$ importance sampling has converged, and there are several criteria available in the literature. The classic effective sample size (ESS)(Kong, Liu, and Wong, 1994)

$$\mathrm{ESS} = \frac{1}{\sum_{i=1}^{N} W_i^2} \in [1, N]$$

arises from an analysis of the asymptotic efficiency of importance sampling estimates: Consider additional $Y^1, \dots, Y^N \overset{\text{i.i.d}}{\sim} \mathbf{P}$, a test function $f \in L^2(\mathbf{P})$ and assume that $\rho < \infty$. We may then estimate $\zeta = \mathbf{P}f$ in two ways: either by using the importance sampling estimate

$$\hat{\zeta}_{\mathrm{IS}} = \hat{\mathbf{P}}_N(f) = \sum_{i=1}^{N} W_i f(X^i) = \frac{1}{N}\sum_{i=1}^{N} (NW_i) f(X^i),$$

or by standard Monte-Carlo integration

$$\hat{\zeta}_{\mathrm{MC}} = \frac{1}{N}\sum_{i=1}^{N} f(Y^i).$$

(Kong, 1992) applies the delta method to $\mathrm{Var}\left(\hat{\zeta}_{\mathrm{IS}}\right)$, obtaining

$$\mathrm{Var}\left(\hat{\zeta}_{\mathrm{IS}}\right) \approx \mathrm{Var}\left(\hat{\zeta}_{\mathrm{MC}}\right)\left(1 + \mathrm{Var}\left(NW_1\right)\right).$$

Note that this approximation does not depend on the specific f considered, and it is not guaranteed that for large N the remainder goes to 0, as (Kong, 1992) mentions. In particular, the approximation has to fail whenever $\mathrm{Var}\left(\hat{\zeta}_{\mathrm{IS}}\right) < \mathrm{Var}\left(\hat{\zeta}_{\mathrm{MC}}\right)$, i.e. when importance sampling actually performs variance reduction. Nevertheless, whenever the approximation is valid, we may interpret

check W / w / tilde w in the following

$$\frac{N}{1 + \mathrm{Var}\left(NW_1\right)}$$

as an effective sample size, in the sense that N times the relative efficiency of $\hat{\zeta}_{\mathrm{MC}}$ relative to $\hat{\zeta}_{\mathrm{IS}}$ is approximately given by this expression. As the self-normalized

replace auto by self

weights $W^1, \dots, W^N$ are exchangeable and sum to 1, their expected value is $\mathbb{E}W_1 = \frac{1}{N}$. Estimating $\mathrm{Var}(W_1)$ by the unadjusted sample covariance $\frac{1}{N}\sum_{i=1}^N W_i^2 - \frac{1}{N^2}$ then results in the promised

$$\mathrm{ESS} = \frac{N}{1 + N^2\left(\frac{1}{N}\sum_{i=1}^N W_i^2 - \frac{1}{N^2}\right)} = \frac{1}{\sum_{i=1}^N W_i^2}.$$

Notice that as the self-normalized weights sum to 1, the ESS is at least 1, as $0 \leq W_i \leq 1$ and at most N by the Cauchy-Schwarz inequality.

If we write the ESS in terms of the unnormalized weights $\tilde{w}$ we see that the efficiency factor (EF) $\mathrm{EF} = \frac{\mathrm{ESS}}{N}$ fulfills

$$\mathrm{EF} = \frac{\mathrm{ESS}}{N} = \frac{\left(\frac{1}{N}\sum_{i=1}^N \tilde{w}_i\right)^2}{\frac{1}{N}\sum_{i=1} \tilde{w}_i^2} \overset{a.s}{\to} \frac{(\mathbf{G}[\tilde{w}])^2}{\mathbf{G}[\tilde{w}^2]} = \rho^{-1},$$

if $\tilde{w} \in L^2(\mathbf{G})$ (Agapiou et al., 2017, Section 2.3.2). Thus, asymptotically, a large ESS leads to small bias and MSE for bounded functions f. Additionally, the above derivations allow us to interpret the second moment

$$\rho = \mathbf{G}[(NW_1)^2] = (\mathbf{G}[NW_1])^2 + \mathrm{Var}(NW_1) = 1 + \mathrm{Var}(NW_1) \approx \frac{\mathrm{Var}\left(\hat{\zeta}_{\mathrm{IS}}\right)}{\mathrm{Var}\left(\hat{\zeta}_{\mathrm{MC}}\right)}$$

as the asymptotic relative efficiency of the two estimators, as long as this approximation is valid. In practice, a small ESS can be an indicator that importance sampling with $\mathbf{G}$ may be inadequate. Note that relying solely on the empirical ESS may lead to problems, see the following example. To prepare, we prove a lemma regarding ρ for Gaussian targets and proposals.

Lemma 3.6. *Let $\mathbf{P} = \mathcal{N}(\mu, \Sigma)$ and $\mathbf{G} = \mathcal{N}(\nu, \Omega)$ be two p-dimensional Gaussian distributions with means $\mu, \nu \in \mathbf{R}^p$ and SPD covariance matrices $\Sigma, \Omega \in \mathbf{R}^{p\times p}$. Then ρ is finite if, and only if, $\Omega \succ \frac{1}{2}\Sigma$.*

Proof. For the weights $w = \frac{p}{g}$ we have

$$\begin{aligned}\rho = \mathbf{G}[w^2] &= \int \frac{p^2(x)}{g^2(x)} g(x)\mathrm{d}x = \int \frac{p^2(x)}{g(x)}\mathrm{d}x \\ &= \int \frac{\sqrt{\det \Omega}}{\sqrt{(2\pi)^p}\det \Sigma} \exp\left(-(x-\mu)^T\Sigma^{-1}(x-\mu) + \frac{1}{2}(x-\nu)^T\Omega^{-1}(x-\nu)\right)\mathrm{d}x.\end{aligned}$$

The exponent is a quadratic form in x, and so the integral is finite if, and only if, the matrix of coefficients, $-\Sigma^{-1} + \frac{1}{2}\Omega^{-1}$ is negative definite. Rearranging terms, we see that this is equivalent to $\Omega \succ \frac{1}{2}\Sigma$. □

Example 3.2 (failure of the ESS)**.** Consider the Gaussian scale mixture

$$\mathbf{P} = \frac{1}{2}\left(\mathcal{N}(0,1) + \mathcal{N}(0, \varepsilon^{-2})\right)$$

and proposal $\mathbf{G} = \mathcal{N}(0,1)$. The weights are then given by

$$w(x) = \frac{1}{2}\left(1 + \frac{\varepsilon}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\left(\varepsilon^2 - 1\right)\right)\right)$$

and their second moment w.r.t. $\mathbf{G}$

$$\rho = \int w^2(x)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)\mathrm{d}x$$

is finite if, and only if, $\varepsilon^2 > \frac{1}{2}$, by the preceding lemma. Thus, for $\varepsilon^2 \leq \frac{1}{2}$ interpreting the ESS or EF is not sensible. Nevertheless, given samples $X^1, \dots, X^N \overset{\text{i.i.d}}{\sim} \mathbf{G}$, we may calculate the ESS in the

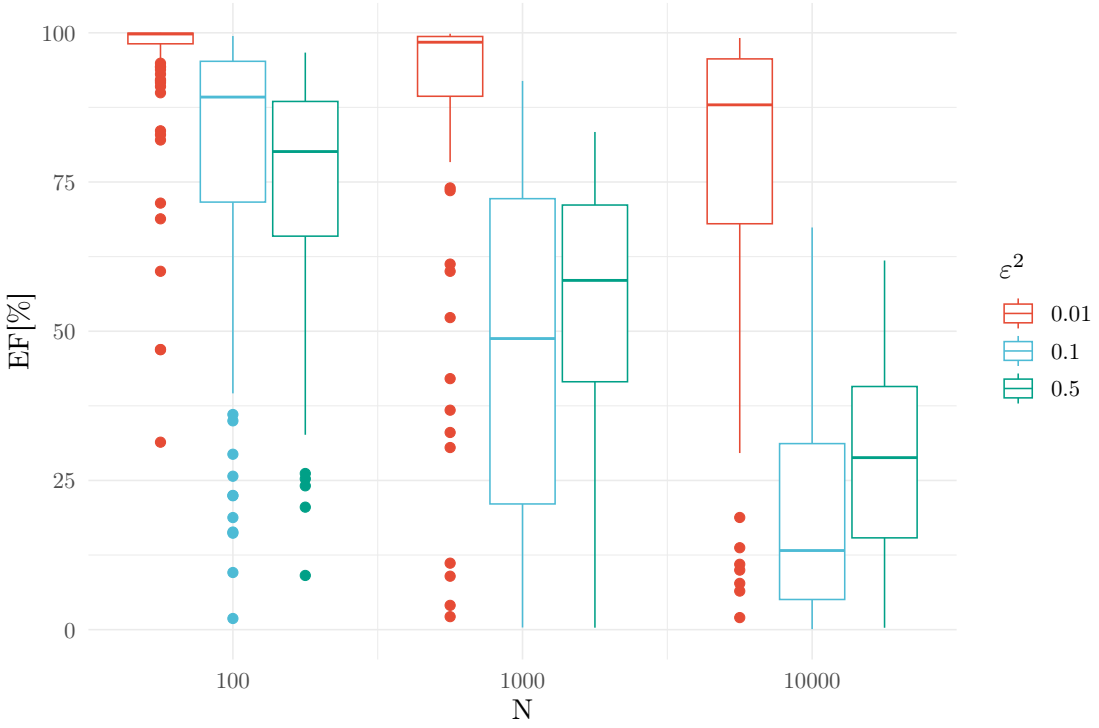

Figure 3.1: Empirical EF for the setup of Example 3.2 for varying sample sizes N and ε^2 and $M = 100$ replications. Here $\mathbf{G} = \mathcal{N}(0,1)$ and $\mathbf{P} = \frac{1}{2}\left(\mathcal{N}(0,1) + \mathcal{N}(0,\varepsilon^{-2})\right)$. In all scenarios the second moment ρ is infinite, thus high EFs are misleading us to believe that importance sampling performs well when it does not.

usual way. If N is only moderately large, there is a high probability that most samples do not lie in a region where weights are small, i.e. in the tails of the second component. Thus, unless N is large, the empirical ESS will be large, deceiving us to think that importance sampling with $\mathbf{G}$ is feasible.

We illustrate this by a simulation study, where we calculate the EF $M = 100$ times for different values of N and ε. We used $N = 100, 1000, 10000$ and $\varepsilon^2 = 0.01, 0.1, 0.5$; the results may be found in Figure 3.1. Notice that for all values of ε considered, we have $\rho = \infty$. We see that even for $N = 1000$ and $\varepsilon = \frac{1}{2}$ the upper quartile of EFs is 71%, which seems reasonable to declare importance sampling to perform well.

As an alternative, we may want to assess whether importance sampling has converged through the empirical variance of $\hat{\zeta}_N$,[2] i.e.,

$$\widehat{\mathrm{Var}}\left(\hat{\zeta}_N\right) = \frac{1}{N}\left(\frac{1}{N}\sum_{i=1}^{N} w_i^2 f(X^i)^2 - \hat{\zeta}_N^2\right)$$

is, while seemingly natural, flawed (Chatterjee and Diaconis, 2018). Indeed, the authors show that for any given threshold ϵ we may find an N which only depends on ϵ, such that the probability that the empirical variance exceeds ϵ for this N is small. This is summarized in the following theorem.

Theorem 3.3 (Chatterjee and Diaconis, 2018, Theorem 2.1)**.** *Given any $\epsilon > 0$, there exists*

lower bound on N?

$N \leq \epsilon^{-2} 2^{1+\epsilon^{-3}}$ such that the following is true. Take any $\mathbf{G}$ and $\mathbf{P}$ as in Theorem 3.2, and any

[2] As the following arguments depend on the sample size N, we mark this dependency by adding N to the subscript of the estimator.

$f : \mathcal{X} \to \mathbf{R}$ such that $\|f\|_{L^2(\mathbf{P})} \leq 1$. Then

$$\mathbb{P}\left(\widehat{\mathrm{Var}}\left(\hat{\zeta}_N\right) < \epsilon\right) \geq 1 - 4\epsilon.$$

make eps / sigma consistent in all examples

The problem here is that N does not depend on $\mathbf{G}$ and $\mathbf{P}$, so we may choose $\mathbf{G}$ almost singular to $\mathbf{P}$. As an example, take $\mathbf{P} = \mathcal{N}(0, 1)$ and $\mathbf{G} = \mathcal{N}(0, \sigma^2)$ for $\sigma^2 > \frac{1}{2}$. The weights are then given by

$$w(x) = \sigma \exp\left(-\frac{x^2}{2}\left(1 - \frac{1}{\sigma^2}\right)\right),$$

and for $X \sim \mathbf{G}$ the variance of $w(X)X$ is

$$\tau^2 = \mathrm{Var}\left(w(X)X\right) = \frac{\sigma^4}{(2\sigma^2 - 1)^{\frac{3}{2}}} \tag{3.10}$$

which goes to ∞ as σ^2 does, see the appendix for the calculations. Thus for a pre-specified $\epsilon > 0$, let N be as in Theorem 3.3 and choose σ^2 such that $\mathrm{Var}\left(\hat{\zeta}_N\right) = \frac{\tau^2}{N}$ is larger than, say, 10ϵ. By the preceding theorem, we would, with large probability, observe a small empirical variance and thus declare $\hat{\zeta}_N$ to have converged, whereas, in reality, we would need a sample size that is 100 times as large.

Thus using the empirical variance as a threshold for convergence should be avoided, at least for importance sampling where the weights can be evaluated exactly. For self-normalized importance sampling, the authors do not provide such a theorem. As a remedy (Chatterjee and Diaconis, 2018) suggest the heuristic $q_N = \mathbb{E}Q_N$ where

$$Q_N = \max_{1 \leq i \leq N} W_i \in [0, 1].$$

This judges whether importance sampling has collapsed to just a few particles and is itself amenable to Monte-Carlo integration, by repeatedly sampling N samples from $\mathbf{G}$ and calculating the weights. As this requires multiple runs of importance sampling, it may, however, be prohibitively expensive in practice.

In the following sections, we will predominantly take the position that we are interested in finding a good particle approximation $\hat{\mathbf{P}}_N$ of the form Equation (3.7) over finding the optimal proposal $\mathbf{G}^*$ Proposition 3.1 and assume that the importance sampling weights can only be evaluated up to a constant. This has several reasons: First of all, for most problems considered in this thesis $\mathbf{P}$ is usually a conditional distribution, e.g. $\mathbf{P} = \mathbb{P}^{X|Y=y}$ for states X and observations Y in the SSM context. Should the appropriate densities exist, evaluating the weights amounts to calculating

$$\frac{\mathrm{d}\mathbb{P}^{X|Y=y}}{\mathrm{d}\mathbf{G}}(x) = \frac{p(x|y)}{g(x)} = \frac{p(y|x)p(x)}{g(x)p(y)} \propto \frac{p(y|x)p(x)}{g(x)}.$$

In these situations $p(y) = \int p(x, y)\,\mathrm{d}x$ is usually intractable. For $\mathbf{G}^*$ we are in the same situation, where the evaluation of the integration constant $\mathbf{P}|f|$ is infeasible, but the density $|f(x)|p(x)$ is available. Second, focusing on the particle approximation allows us to consider multiple test functions f, e.g. focus on different marginals of $\mathbf{P}$, which is usually what practitioners are interested in. Finally, this allows us to simplify the notation used in this thesis. $\mathbf{P}$ will always be the probability measure of interest and $\mathbf{G}$ the proposal. In later parts of this thesis, we will predominantly perform Gaussian importance sampling, i.e. $\mathbf{G} = \mathcal{N}(\mu, \Sigma)$, hence a handy mnemonic is to think of $\mathbf{G}$ as a **G**aussian proposal.

Let us now turn towards the problem of finding a good proposal $\mathbf{G}$ for a given $\mathbf{P}$.

3.4.1 Laplace approximation (LA)

The Laplace approximation (LA) goes back to Laplace (Laplace, 1986) who invented the technique to approximate moments of otherwise intractable distributions. Since (Tierney and Kadane, 1986; Tierney, Kass, and Kadane, 1989) rediscovered its use to approximate posterior means and variances, it has been a staple method for approximate inference. The method is based on a second-order Taylor series expansion of the log target density $\log p(x)$ around its mode $\hat{x}$, i.e. matching mode and curvature. Assuming the density is sufficiently smooth, we have

$$\log p(x) \approx \log p(\hat{x}) + \underbrace{\nabla_x \log p(\hat{x})}_{=0}(x - \hat{x}) + \frac{1}{2}(x - \hat{x})^T H (x - \hat{x}) \tag{3.11}$$

where H is the Hessian of $\log p$ evaluated at $\hat{x}$. As $\log p(\hat{x})$ does not depend on x, the right-hand side can be seen (up to additive constants) as the density of a Gaussian distribution with mean $\hat{x}$ and covariance matrix $\Sigma = -H^{-1}$. Thus using $\mathbf{G} = \mathcal{N}(\hat{x}, -H^{-1})$ as a proposal in importance sampling seems promising. If $\hat{x}$ is the unique global mode of p and H is negative definite, the LA yields an actual Gaussian distribution. To obtain the LA in practice, a Newton-Raphson scheme may be used, which conveniently tracks H as well. Furthermore, if $\mathbf{P}$ includes more structure, e.g. it is the smoothing density in the SSM context, we may be able to exploit this structure to design efficient Newton-Raphson schemes, see Section 3.6.1.

The main advantage of the LA is that it is usually fast to obtain and, for sufficiently well-behaved distributions on a moderate dimensional space, provides reasonably high ESS. Additionally, the Newton-Raphson iterations to find the mode and Hessian are robust and require no simulation, unlike the other methods discussed further below. For the SSMs we consider in this thesis, the numerical methods can be implemented using the Kalman filter and smoother (Durbin and Koopman, 1997; Shephard and Pitt, 1997), even in the degenerate case where H is indefinite (Jungbacker and Koopman, 2007), see also Section 3.6.1.

more theoretical background on LA?

However, as the LA is a local approximation, it may be an inappropriate description of the global behavior of the target, see Example 3.3 for a breakdown of LA, and the simulation studies presented in Section 3.8. Additionally, even if the LA works in principle, its ESS will usually degenerate quickly once the dimension increases whereas the Cross-Entropy method (CE-method) and Efficient Importance Sampling (EIS) do so at a slower pace.

3.4.2 The Cross-Entropy method (CE-method)

Recall from our discussion surrounding Theorem 3.2 that for importance sampling to be effective, a small KL-divergence between the target $\mathbf{P}$ and the proposal $\mathbf{G}$ implies good performance for importance sampling. As the KL-divergence depends on global properties of $\mathbf{P}$, i.e. the Radon-Nikodym derivative $\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{G}}$, minimizing it leads to a global approximation of $\mathbf{P}$, improving on the local-approximation provided by the LA.

The Cross-Entropy method (CE-method) (Rubinstein, 1999; Rubinstein and Kroese, 2004) implements this idea and selects from a parametric family $(\mathbf{G}_\psi)_{\psi\in\Psi}$ of proposals the one that minimizes the Kullback Leibler divergence (KL-divergence) to the target. Here Ψ is usually a subset of $\mathbf{R}^k$, which may be open, closed or neither. Thus, the CE-method aims at solving the following optimization problem

$$\min_{\psi\in\Psi} \mathcal{D}_{\mathrm{KL}}\left(\mathbf{P} || \mathbf{G}_\psi\right),$$

for the optimal ψ_{CE}, should the minimum exist. The existence and uniqueness of ψ_{CE} will depend heavily on the choice of parametric family $(\mathbf{G}_\psi)_{\psi\in\Psi}$ and $\mathbf{P}$.

We will assume the existence of a common dominating measure μ for both $\mathbf{P}$ and all $\mathbf{G}_\psi$, $\psi \in \Psi$ with corresponding densities p and g_ψ, $\psi \in \Psi$. The importance sampling weights are then given by

$$w_\psi(x) = \frac{p(x)}{g_\psi(x)},$$

$x \in \mathcal{X}$, or, if at least one of p and g_ψ is only available up to a constant, by

$$\tilde{w}_\psi(x) \propto \frac{p(x)}{g_\psi(x)}.$$

If the dependence on ψ is not of interest or the particular ψ is obvious from the context, we may drop the subscript.

The KL-divergence is given by

$$\mathcal{D}_{\mathrm{KL}}\left(\mathbf{P}||\mathbf{G}_\psi\right) = \mathbf{P}\left[\log w_\psi\right],$$

and can be infinite, e.g. if $\mathbf{P}$ does not possess second moments and $\mathbf{G}_\psi$ are Gaussian distributions. If the KL-divergence is infinite for all $\psi \in \Psi$, the CE-method becomes uninteresting. As such we will require that the KL-divergence is finite for at least one $\psi \in \Psi$, and restrict Ψ, without loss of generality, to those ψ where the KL-divergence is finite.

As the appropriate densities exist, we may reformulate the optimization problem to maximize the cross-entropy between p and g_ψ instead:

$$\begin{aligned}\operatorname{argmin}_{\psi\in\Psi} \mathcal{D}_{\mathrm{KL}}\left(\mathbf{P}||\mathbf{G}_\psi\right) &= \operatorname{argmin}_{\psi\in\Psi} \mathbf{P}\left[\log p\right] - \mathbf{P}\left[\log g_\psi\right] \\ &= \operatorname{argmax}_{\psi\in\Psi} \mathbf{P}\left[\log g_\psi\right]\end{aligned}. \tag{3.12}$$

As the KL-divergence is non-negative by the information inequality, the cross-entropy $\mathbf{P}\left[\log g_\psi\right]$ is bounded from above by the differential entropy of $\mathbf{P}$, $\mathbf{P}[\log p]$. For centered distributions with covariance matrix Σ the differential entropy is bounded above by the maximum entropy distribution in this setting, the Gaussian $\mathcal{N}(0, \Sigma)$ (Cover and Thomas, 2006, Example 12.2.8). Thus, if second moments of $\mathbf{P}$ exist, the cross-entropy is bounded from above, and so a maximizer exists if the supremum over Ψ is attained. This would be the case if Ψ is compact and $\psi \mapsto \mathbf{P}[\log g_\psi]$ is continuous, however compact Ψ is too restrictive for our purposes. Instead, we are going to focus on more realistic assumptions.

Suppose now that $\psi \mapsto \log g_\psi(x)$ is (strictly) concave for $\mathbf{P}$-almost every $x \in \mathcal{X}$ and Ψ is a convex subset of $\mathbf{R}^k$. Then $\psi \mapsto \mathbf{P}\left[\log g_\psi\right]$ is (strictly) concave as well. As a consequence, we may apply the usual results from convex optimization, i.e. every local maximum is a global one and if $\psi \mapsto \log g_\psi(x)$ is strictly convex for $\mathbf{P}$-almost every x, there is at most one maximizer (Bazaraa, Sherali, and Shetty, 2006, Theorem 3.4.2).

As we have seen in Lemma 3.3, the densities of exponential families are log-concave in the natural parameter, and as such they will be the primary candidates for our investigations of the CE-method. If we use proposals from an exponential family, we may get rid of the base measure term $h(x)$ in the densities, as the following lemma shows.

Lemma 3.7. *Let $\mathbf{P}$ be a probability measure on $\mathcal{X} = \mathbf{R}^p$ and let $(\mathbf{G}_\psi)_{\psi\in\Psi}$ be a natural exponential family on $\mathcal{X}$ such that $\mathbf{P} \ll \mathbf{G}_\psi$ for all $\psi \in \Psi$. Let μ be the dominating measure of the exponential family, such that*

$$\frac{\mathrm{d}\mathbf{G}_\psi}{\mathrm{d}\mu}(x) = \frac{h(x)}{Z(\psi)} \exp\left(\psi^T T(x)\right),$$

with $h \geq 0$ μ-a.s.

Then $h\mu$ is a dominating measure for both $\mathbf{P}$ and $\mathbf{G}_\psi$ for every ψ in Ψ.

Proof. Let $A \subseteq \mathbf{R}^p$ be measurable. As h is a.s. non-negative, $(h\mu)(A) = 0$ implies that $h\mathbf{1}_A = 0$ μ-a.s. Thus $\mathbf{G}_\psi(A) = \int \mathbf{1}_A(x) \frac{h(x)}{Z(\psi)} \exp(\psi^T T(x)) \,\mathrm{d}\mu = 0$ for all ψ as well. As $\mathbf{G}_\psi \gg \mathbf{P}$ and $\gg$ is transitive, $h\mu$ dominates $\mathbf{P}$ as well. □

As a consequence, when performing importance sampling with target $\mathbf{P}$ and proposal $\mathbf{G}_\psi$ from an exponential family, we will assume in the following that $h \equiv 1$, achieved by taking $h\mu$ as the joint dominating measure.

An additional attractive property of the CE-method for exponential families with natural parameter $\psi \in \mathbf{R}^k$ is that the optimal ψ_{CE} only depends on the expected value $\mathbf{P}[T]$. We first show, that if the covariance of the sufficient statistic is positive definite, the expected value of T under $\mathbf{G}_\psi$ uniquely determines $\psi \in \Psi$, see also (Brown, 1986, Corollary 2.5) for a similar result in minimal exponential families.

Lemma 3.8. *Let $(\mathbf{G}_\psi)_{\psi\in\Psi}$ form a k-dimensional natural exponential family with log-densities*

$$\log g_\psi(x) = \psi^T T(x) - \log Z(\psi),$$

and convex parameter space $\Psi \subseteq \mathbf{R}^k$. Let $\psi, \psi' \in \operatorname{int}\Psi$ with $\mathbf{G}_\psi[T] = \mathbf{G}_{\psi'}[T]$. If $\operatorname{Cov}_{\mathbf{G}_\psi} T$ is positive definite, then ψ and ψ' coincide.

Proof. Consider the function $b : \Psi \to [-\infty, \infty)$

$$\xi \mapsto b(\xi) = \mathbf{G}_{\psi'}[\log g_\xi] = \xi^T \mathbf{G}_{\psi'}[T] - \log Z(\xi).$$

By Theorem 3.1, $\mathbf{G}_{\psi'}[T]$ is finite and b possesses derivatives of every order. Then ψ is a critical point of this map, as the gradient at ψ is

$$\mathbf{G}_{\psi'}[T] - \nabla_\psi \log Z(\psi) = \mathbf{G}_{\psi'}[T] - \mathbf{G}_\psi[T] = 0.$$

The Hessian of this function at ξ is, see Theorem 3.1,

$$-H_\xi \log Z(\xi) = -\operatorname{Cov}_{\mathbf{G}_\xi}[T],$$

which is negative semi-definite, so b is concave. At $\xi = \psi$ it is even negative definite, so the critical point ψ is a strict local maximum. By concavity, it is the unique global maximum, and thus the unique critical point, so $\psi = \psi'$. □

Proposition 3.2 (The CE-method for exponential families)**.** *Let $(\mathbf{G}_\psi)_{\psi\in\Psi}$ form a k-dimensional natural exponential family with log-densities*

$$\log g_\psi(x) = \psi^T T(x) - \log Z(\psi),$$

and convex parameter space $\Psi \subseteq \mathbf{R}^k$. Suppose $T \in L^1(\mathbf{P})$.

If there is a $\psi_{CE} \in \Psi$ such that

$$\mathbf{P}[T] = \mathbf{G}_{\psi_{CE}}[T],$$

then ψ_{CE} is a maximizer of Equation (3.12). Furthermore, if $\operatorname{Cov}_{\mathbf{G}_{\psi_{CE}}} T$ is positive definite the maximizer is unique.

Proof. The target may be rewritten as

$$\psi \mapsto f(\psi) = \mathbf{P}\left[\log g_\psi(x)\right] = \log Z(\psi) + \psi^T \mathbf{P}[T].$$

As $\log Z(\psi)$ is the cumulant-generating function of $\mathbf{G}_\psi$ it is twice differentiable, and so is f. The gradient of $\log Z(\psi)$ is

$$\nabla_\psi \log Z(\psi) = \mathbf{G}_\psi[T]$$

and its Hessian is

$$H_\psi \log Z(\psi) = \operatorname{Cov}_{\mathbf{G}_\psi}(T)$$

the covariance of T under $\mathbf{G}_\psi$. Thus the Hessian of f is

$$H_\psi f = -\operatorname{Cov}_{\mathbf{G}_\psi}(T),$$

which is negative-semi-definite. Therefore f is concave, and any local maximizer ψ is a global maximizer. The gradient of f is

$$\nabla_\psi f(\psi) = \mathbf{P}[T] - \mathbf{G}_\psi[T],$$

which is equal to 0 if, and only if, ψ solves

$$\mathbf{P}[T] = \mathbf{G}_\psi[T].$$

Uniqueness follows from the preceding Lemma 3.8. □

As a consequence, the CE-method for natural exponential families reduces to matching the moments of the sufficient statistic of the target and proposal. In many cases, this system of equations can be solved analytically or by gradient descent algorithms. Let us discuss the assumptions and applicability of this proposition. Assuming that $T \in L^1(\mathbf{P})$ is necessary for the target to be finite, it cannot be dropped. As T typically consists of polynomial, rational or exponential functions, this is not too restrictive, provided the target does not exhibit heavy tails. The proof of uniqueness relies on $\mathrm{Cov}_{\mathbf{G}_\psi} T$ being positive definite, to ensure that $\psi \mapsto \log Z(\psi)$ is strictly convex. This could also be achieved by requiring the exponential family to be minimal, see (Brown, 1986, Theorem 1.13 (iv)). The existence of a ψ such that $\mathbf{P}[T] = \mathbf{G}_\psi[T]$ is not restrictive for most commonly used distributions: for the (multivariate) normal, Poisson, negative binomial and binomial distribution there is always a unique solution, as the sufficient statistics consist of means and covariances.

While $\mathbf{P}[T]$ is usually not available, it is itself amenable to importance sampling. Given a proposal $\mathbf{G}$ we may estimate $\mathbf{P}[T]$ by $\hat{\mathbf{P}}_N T = \sum_{i=1}^N W^i T(X^i)$ for $X^1, \dots, X^N \overset{\text{i.i.d}}{\sim} \mathbf{G}$ and auto-normalized importance sampling weights W^i and in turn, applying Proposition 3.2, estimate ψ_{CE} by $\hat{\psi}_{\mathrm{CE}}$ solving

$$\hat{\mathbf{P}}_N[T] = \mathbf{G}_{\hat{\psi}_{\mathrm{CE}}}[T]. \tag{3.13}$$

As $T \in L^1(\hat{\mathbf{P}}_N)$, the only conditions we have to check to apply the above proposition are that this equation has a unique solution $\mathbf{G}$-almost surely in the interior of Ψ and that Ψ is convex.

To apply the CE-method in practice, one usually iterates the sampling and estimation steps, using the previously found $\hat{\psi}_{\mathrm{CE}}$ to sample in the current iteration and starting the iteration with a proposal from the same exponential family $\mathbf{G} = \mathbf{G}_{\psi^0}$. To ensure numerical convergence, a popular device is that of common random numbers (CRNs), i.e. using the same random number seed in all iterations. A basic version of the CE-method is presented in Algorithm 4.

Algorithm 4 The basic CE-method algorithm for exponential families

Require: exponential family $(\mathbf{G}_\psi)_{\psi \in \Psi}$, initial ψ^0, sample size N, unnormalized weights $\tilde{w}$

set $l = 0$
store random number seed
repeat
  restore random number seed
  sample $X^1, \dots, X^N \sim \mathbf{G}_{\psi^l}$
  calculate self-normalized weights W^i for $i = 1, \dots, N$
  estimate $\hat{\psi}_{\mathrm{CE}}$ ▷ Equation (3.13)
  set $\psi^{l+1} = \hat{\psi}_{\mathrm{CE}}$
  set $l = l + 1$
until $\hat{\psi}^l$ converged
return $\hat{\psi}_{\mathrm{CE}} = \hat{\psi}^l$

literature review CEM

The CE-method is routinely used for estimating failure probabilities for rare events (Homem-de-Mello, 2007) and has been applied to Bayesian posterior inference (Ehre et al., 2023; Engel et al., 2023), Bayesian marginal likelihood estimation (J. C. C. Chan and Eisenstat, 2012) and optimal control problems (Kappen and Ruiz, 2016; Zhang et al., 2014).

more lit. review CEM

Importance sampling is well known to exhibit the curse of dimensionality (COD) Bengtsson, Bickel, and B. Li, 2008, i.e. the phenomenon that in many problems, unless N grows exponentially with the dimension of $\mathcal{X}$, the weights collapse to a single particle, i.e. $W^{(N)} \to 1$ as the dimension of $\mathcal{X}$ goes to ∞. As the CE-method employs importance sampling to obtain $\hat{\psi}_{\mathrm{CE}}$, it too is affected by this phenomenon, see also Section 3.8. The screening method Rubinstein and Glynn, 2009 deals with the COD by keeping components of ψ^l that vary too much from iteration to iteration fixed, in essence reducing the dimension of Ψ. Alternatively, the improved cross-entropy method (J. C. C.

Chan and Kroese, 2012) suggests generating approximately independent samples from $\mathbf{P}$ by, e.g., MCMC-methods, and replacing the importance sampling version of $\hat{\mathbf{P}}_N$ in Equation (3.13) by the actual empirical distribution. Still, in high dimensions both of these approaches may be difficult to implement: the screening method may not move far from the initial proposal and MCMC-methods are expensive in high dimensions.

As stated in (J. C. C. Chan and Kroese, 2012) there may be two reasons as to why the CE-method fails: either the parametric family is not rich enough to give a good approximation to $\mathbf{P}$, i.e. $\mathcal{D}_{\mathrm{KL}}(\mathbf{P}||\mathbf{G}_{\psi_{\mathrm{CE}}})$ is still large, or the estimate $\hat{\psi}_{\mathrm{CE}}$ fails to be close to ψ_{CE}. As our simulation studies Section 3.8 suggest, the reason for the degeneracy seems to be the latter. It will thus be beneficial to investigate the asymptotic behavior of $\hat{\psi}_{\mathrm{CE}}$.

In the remainder of this section, we will derive novel results on the performance of the estimator $\hat{\psi}_{\mathrm{CE}}$ of ψ_{CE}. In particular, we will investigate under which conditions $\hat{\psi}_{\mathrm{CE}}$ is consistent and asymptotically normal. To focus on the asymptotic behavior, we will only perform a single iteration of the basic CE-method algorithm (Algorithm 4). While we restrict ourselves here to the setting of k-dimensional natural exponential families, these results should generalize to other classes of distributions as well. The advantage that this class of families has is that due to the structure of the densities, they provide straightforward (regularity) conditions for the asymptotic results to hold. As the target functions are concave, these conditions are rather liberal. We start with proving the consistency of $\hat{\psi}_{\mathrm{CE}}$.

Theorem 3.4 (consistency of $\hat{\psi}_{\mathrm{CE}}$)**.** *Adopt the same assumptions as in Proposition 3.2. Furthermore, let $\mathbf{G} \gg \mathbf{P}$ be a proposal distribution and assume that*

(i) ψ_{CE} is the unique maximizer of Equation (3.12),

(ii) ψ_{CE} is in the interior of the convex parameter space Ψ.

Then $\hat{\psi}_{CE}$ is a strongly consistent estimator of ψ_{CE}.

The proof is based on the following theorem of Haberman.

Theorem 3.5 ((Haberman, 1989, Theorem 5.1)[3])**.** *Let $\Psi \subseteq \mathbf{R}^k$, $\mathcal{X}$ a separable, complete metric space and $b_{\mathcal{X}} : \mathcal{X} \times \mathbf{R}^k \to [-\infty, \infty)$ such that for every $x \in \mathcal{X}$ the function*

$$b(x, \cdot) : \mathbf{R}^k \to [-\infty, \infty), \psi \mapsto b(x, \psi)$$

is concave. Let $\mathbf{P}$ be a probability measure on $\mathcal{X}$ such that $\mathbf{P}[b(\cdot, \psi)] < \infty$ for all $\psi \in \mathbf{R}^k$. Assume that $\psi^* \in \Psi$ is the unique maximizer of

$$b_{\Psi} : \Psi \to [-\infty, \infty), \psi \mapsto \mathbf{P}[b(\cdot, \psi)].$$

Let $(X^i)_{i \in \mathbf{N}} \overset{i.i.d}{\sim} \mathbf{P}$ be a sequence of i.i.d. random variables with distribution $\mathbf{P}$ and let for $N \in \mathbf{N}$ let

$$\hat{\mathbf{P}}_N = \frac{1}{N} \sum_{i=1}^{N} \delta_{X^i}$$

be their empirical distribution. Let $\left(\hat{\psi}_N\right)_{N \in \mathbf{N}}$ be a sequence of M-estimators, i.e. a sequence of maximizers of

$$\hat{b}_{\Psi} : \Psi \to [-\infty, \infty), \psi \mapsto \hat{\mathbf{P}}_N[b(\cdot, \psi)].$$

Assume that the following conditions hold:

(C1) For some closed set V, ψ^* is in the interior of V and $\Psi \cap V$ is closed.

(C2) ψ^* is the unique maximizer of

$$b_{\mathrm{cl}(\Psi)} : \mathrm{cl}(\Psi) \to [-\infty, \infty), \psi \mapsto \mathbf{P}[b(\cdot, \psi)],$$

where cl *denotes the closure of Ψ in $\mathbf{R}^k$.*

[3]Note that while the actual theorem assumes conditions 1,2,5 and 6 in the paper, C3 as stated here implies conditions 5 and 6, see also the discussion in Sections 2.3 and 2.4 in (Haberman, 1989).

(C3) Ψ *is convex and* b_Ψ *is finite on a nonempty open set.*

Then

$$\hat{\psi}_N \overset{N\to\infty}{\longrightarrow} \psi^*$$

$\mathbf{P}$*-almost surely, so* $\hat{\psi}_N$ *is strongly consistent.*

The assumptions of this theorem ensure that the unique optimum is in the interior of Ψ and „well-separated“ from its boundary, so there are no additional maximizers on the boundary. In this case, concavity of $b(x,\psi)$ together with the law of large numbers yield uniform convergence of $\hat{\mathbf{P}}_N[b(\cdot,\psi)] \to \mathbf{P}[b(\cdot,\psi)]$ on compacta and thus also for $\hat{\psi}_N$, see (Haberman, 1989, pp. 1652).

To apply this theorem to our setting, let us begin by extending it to incorporate importance sampling.

Proposition 3.3. *Assume that the conditions of Theorem 3.5 are fulfilled and let* $\mathbf{G} \gg \mathbf{P}$ *be another probability measure with Radon-Nikdoym derivative* $w(x) = \frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{G}}(x)$*. Let* $(X^i)_{i\in\mathbf{N}} \overset{i.i.d}{\sim} \mathbf{P}$ *and consider the particle approximations*

$$\tilde{\mathbf{P}}_N = \frac{1}{N}\sum_{i=1}^{N} w(X^i)\delta_{X^i},$$
$$\hat{\mathbf{P}}_N = \sum_{i=1}^{N} W^i \delta_{X^i},$$

and suppose for every $N \in \mathbf{N}$ *there exist M-estimators*

$$\tilde{\psi}_N \in \operatorname{argmax}_{\psi\in\Psi} \tilde{\mathbf{P}}_N\left[b(\cdot,\psi)\right],$$
$$\hat{\psi}_N \in \operatorname{argmax}_{\psi\in\Psi} \hat{\mathbf{P}}_N\left[b(\cdot,\psi)\right].$$

Then both $\tilde{\psi}_N$ *and* $\hat{\psi}_N$ *are strongly consistent estimators of* ψ^**.*

Proof. Define a new objective function $\tilde{b} : \mathcal{X}\times\mathbf{R}^k \to [-\infty,\infty)$ by

$$\tilde{b}(x,\psi) = w(x)b(x,\psi).$$

Then $\mathbf{G}\left[\tilde{b}(\cdot,\psi)\right] = \mathbf{P}\left[b(\cdot,\psi)\right]$ for all $\psi\in\Psi$, and so ψ^* is the unique global maximum of

$$\psi \mapsto \mathbf{G}\left[\tilde{b}(\cdot,\psi)\right].$$

As $\mathbf{G}[\tilde{b}(\cdot,\psi)] = \mathbf{P}[b(\cdot,\psi)] < \infty$ and for fixed $x\in\mathcal{X}$ $\tilde{b}(x,\cdot) = w(x)b(x,\cdot)$ is concave, we may directly apply Theorem 3.5 to $\tilde{\psi}_N$, showing its strong consistency.

For $\hat{\psi}_N$, notice that for a fixed sample $X^1,\dots,X^N \overset{\text{i.i.d}}{\sim} \mathbf{G}$ and any function $f:\mathcal{X}\to[-\infty,\infty)$ we have, a.s.,

$$\hat{\mathbf{P}}_N[f] = \sum_{i=1}^{N} W^i f(X^i) = \frac{\mathbf{G}[\tilde{w}]}{\sum_{i=1}^N \tilde{w}(X^i)} \sum_{i=1}^{N} \frac{\tilde{w}(X^i)}{\mathbf{G}[\tilde{w}]} f(X^i) = \frac{\mathbf{G}[\tilde{w}]}{\sum_{i=1}^N \tilde{w}(X^i)} \tilde{\mathbf{P}}_N[f] \propto \tilde{\mathbf{P}}_N[f],$$

where $\tilde{w}$ are the unnormalized weights, i.e. $\frac{\tilde{w}(x)}{\mathbf{G}[\tilde{w}]} = w(x), x\in\mathcal{X}$. Thus $\hat{\psi}_N$ maximizes $\tilde{\mathbf{P}}_N[b(\cdot,\psi)]$ as well, and the result follows from the consistency of $\tilde{\psi}_N$. □

Let us now prove the promised consistency of the CE-method.

Proof (Theorem 3.4). We show that the assumptions of Theorem 3.5 are fulfilled. Let

$$b : \mathbf{R}^p \times \mathbf{R}^k \to [-\infty, \infty) \qquad b(x, \psi) = \begin{cases} \log g_\psi(x) & \psi \in \Psi, \\ -\infty & \text{else.} \end{cases}$$

As Ψ is convex and $g_\psi(x)$ is log-concave (see Lemma 3.3), $b(x, \cdot)$ is concave. Let $X^1, \dots, X^N \overset{\text{i.i.d}}{\sim} \mathbf{P}$ and let $\tilde{\mathbf{P}}_N = \frac{1}{N} \sum_{i=1}^N \delta_{X^i}$. For $\psi \in \Psi$ we have

$$\mathbf{P}[b(\cdot, \psi)] = \psi^T \mathbf{P}[T] - \log Z(\psi) < \infty,$$

as $T \in L^1(\mathbf{P})$, while for $\psi \notin \Psi$ this integral is $-\infty$. Thus we only have to check that (C1)-(C3) are fulfilled.

For condition (C1) note that, as ψ_{CE} is in the interior of Ψ, we may choose $\varepsilon > 0$ such that the closed ε ball around ψ_{CE}, $\bar{B}_\varepsilon(\psi_{\text{CE}})$ is completely contained in Ψ, so letting $V = \bar{B}_\varepsilon(\psi_{\text{CE}})$ implies the condition. condition (C2) is fulfilled by the definition of b and condition (C3) is fulfilled by considering the neighborhood of ψ_{CE} that is assumed to be contained in Ψ. Finally, by Proposition 3.3, $\hat{\psi}_{\text{CE}}$ is strongly consistent. □

The assumptions on ψ_{CE} and Ψ in Theorem 3.4 could be somewhat looser, as the concavity of the target function is a rather strong property. In natural exponential families,

$$\Psi = \left\{ \psi \in \mathbf{R}^k \,:\, Z(\psi) < \infty \right\}$$

is always convex so this is not a strong restriction. In regular exponential families, Ψ is open and so only the existence and uniqueness of ψ_{CE} are required. Uniqueness may be attained, e.g., by Lemma 3.8. It will also hold if the exponential family considered is minimal (Brown, 1986, Corollary 2.5). Existence is a matter of correctly specifying the exponential family. For example, in Section 3.6.2 we will exploit the Markov structure of targets to restrict ourselves to Gaussian Markov processes for $(\mathbf{G}_\psi)_{\psi \in \Psi}$.

Not only is $\log g_\psi$ concave, but it also possesses derivatives of any order, at least on the interior of Ψ. Indeed, its Hessian is given by the inverse of the Fisher-information matrix $I(\psi)^{-1}$:

$$H_\psi \log g_\psi = -H_\psi \log Z(\psi) = -\operatorname{Cov}_{\mathbf{G}_\psi}(T) = -I(\psi)^{-1}.$$

These rather strong properties enable us to derive a central limit theorem for the CE-method with natural exponential family proposals under quite liberal conditions.

Theorem 3.6 (CLT for $\hat{\psi}_{\text{CE}}$)**.** *Adopt the same assumptions as in Proposition 3.2. Furthermore, let $\mathbf{G} \gg \mathbf{P}$ be a proposal distribution with weights $w = \frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{G}}$ and assume that*

(i) $\psi_{CE} \in \Psi$ is the unique maximizer of Equation (3.12) which lies in the interior of the convex parameter space Ψ,

(ii) the Fisher information matrix $I(\psi_{CE})$ exists and is positive definite,

(iii) $w, wT \in L^2(\mathbf{G})$, and

(iv) $T \in L^2(\mathbf{P})$.

Then

$$\sqrt{N}(\hat{\psi}_{CE} - \psi_{CE}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, BMB)$$

where $B = I(\psi_{CE}) = \operatorname{Cov}_{\mathbf{G}_{\psi_{CE}}}(T)^{-1}$ and

$$M = \operatorname{Cov}_{\mathbf{G}}(wT) = \mathbf{G}\left[w^2 (T - \mathbf{P}[T])(T - \mathbf{P}[T])^T\right] = \mathbf{P}\left[w (T - \mathbf{P}[T])(T - \mathbf{P}[T])^T\right].$$

To prove Theorem 3.6, let us start again with a general version of a central limit theorem for M-estimators based on concave objective functions.

Theorem 3.7 ((Haberman, 1989, Theorem 6.1)[4]). *Consider the same setting as in Theorem 3.5. Assume further that ψ^* lies in the interior of Ψ and that the following conditions hold:*

(C7) The Hessian $H_\psi \mathbf{P}\left[b(\cdot, \psi^*)\right]$ exists and is non-singular.

(C10) For $X \sim \mathbf{P}$ and some neighborhood V of ψ^*

$$\sigma^2(\psi, \xi) = \mathbb{E}\left(b'(X, \psi, \xi)\right)^2 < \infty \qquad \psi \in V, \xi \in \mathbf{R}^k,$$

where $b'(x, \psi, \xi) = \lim_{a \downarrow} a^{-1}\left(b(x, \psi + a\xi) - b(x, \psi)\right)$ is the directional derivative. Note that if b is differentiable for all $\psi \in V$, $b'(x, \psi, \xi) = \xi^T \nabla_\psi b(x, \psi)$ and it suffices to assume $(\nabla_\psi b(x, \psi))_i (\nabla_\psi b(x, \psi))_j \in L^1(\mathbf{P})$ for all $\psi \in N$ and $i, j = 1, \ldots, k$.

Let $M = \operatorname{Cov}\left(\nabla_\psi b(X, \psi)\right)$ and let $B = -\left(H_\psi \mathbf{P}\left[b(\cdot, \psi)\right]\right)^{-1}$. Then

$$\sqrt{N}\left(\hat{\psi}_N - \psi\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, BMB). \tag{3.14}$$

Similar to the consistency result above (Proposition 3.3), we need to extend this CLT to account for importance sampling.

Proposition 3.4. *Assume that the conditions of Theorem 3.7 are fulfilled and use the same notation as in Proposition 3.3. Furthermore, assume that*

(i) $w(\cdot) b'(\cdot, \psi, \xi) \in L^2(\mathbf{G})$ in a neighborhood N of ψ^* for all $\xi \in \mathbf{R}^k$.

rewrite differentiable + second moment

Then

$$\sqrt{N}\left(\tilde{\psi}_N - \psi^*\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, BMB), \tag{3.15}$$

where $M = \operatorname{Cov}\left(w(X) \nabla_\psi b(X, \psi^*)\right)$ for $X \sim \mathbf{G}$ and $B = -\left(H_\psi \mathbf{P}\left[b(\cdot, \psi^*)\right]\right)^{-1}$ is as in Theorem 3.7. Additionally

$$\sqrt{N}\left(\hat{\psi}_N - \psi^*\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, BMB). \tag{3.16}$$

Proof. Similar to the proof of Proposition 3.3, define the new objective function $\tilde{b} : \mathcal{X} \times \mathbf{R}^k \to [-\infty, \infty)$ by

$$\tilde{b}(x, \psi) = w(x) b(x, \psi),$$

and notice that $\mathbf{G}\left[\tilde{b}(\cdot, \psi)\right] = \mathbf{P}\left[b(\cdot, \psi)\right]$. Let us verify the conditions of Theorem 3.7 for $\tilde{b}$ and the probability measure $\mathbf{G}$.

For condition (C7), as $H_\psi \mathbf{P}[b(\cdot, \psi)]$ exists and is non-singular, so does

$$H_\psi \mathbf{G}[\tilde{b}(\cdot, \psi)] = H_\psi \mathbf{P}[b(\cdot, \psi)]$$

exist and is non-singular. Similarly, it is easy to see that $\tilde{b}'(x, \psi, \xi) = w(x) b'(x, \psi, \xi)$ and so for $X \sim \mathbf{G}$

$$\sigma^2_{\tilde{b}}(\psi, \xi) = \mathbb{E}\left(\tilde{b}'(X, \psi, \xi)\right)^2 = \mathbb{E} w^2(X) b'(X, \psi, \xi)^2 < \infty$$

by assumption (i), showing condition (C10). Thus we may apply Theorem 3.7 to $\tilde{b}$ and $\mathbf{G}$, finishing the proof. □

Interestingly, importance sampling only affects the M component of the asymptotic variance. The reason for this is that M is a quadratic function of the weights w, while B only depends linearly on w, allowing to switch integrators from $\mathbf{G}$ to $\mathbf{P}$. We now have all the tools at our disposal to proof Theorem 3.6.

[4] Note, again, that the original theorem is based on conditions 7,8,9 in the paper. However, under (C7), condition (C10) implies conditions 8 and 9 in the paper. See the discussion in Section 3.1 in (Haberman, 1989).

Proof of Theorem 3.6. We show that the assumptions and conditions of Theorem 3.7 for the objective function $b : \mathcal{X} \times \mathbf{R}^k \to [-\infty, \infty)$

$$b(x, \psi) = \begin{cases} \log g_\psi(x) & x \in \Psi \\ -\infty & \text{else,} \end{cases}$$

are fulfilled, which, together with Proposition 3.4 will show the claim.

The Hessian of the objective function is, for $\psi \in \operatorname{int} \Psi$

$$H_\psi \mathbf{P}\left[b(\cdot, \psi)\right] = H_\psi \mathbf{P}\left[\psi^T T - \log Z(\psi)\right] = -H_\psi \log Z(\psi) = -I(\psi),$$

as the cumulant generating function is smooth on $\operatorname{int} \Psi$ (Theorem 3.1). Thus the Hessian is non-singular by assumption (ii), showing that condition (C7) is fulfilled.

For condition (C10), note that for $\psi \in \operatorname{int} \Psi$, b is differentiable with gradient

$$\nabla_\psi b(x, \psi) = T(x) - \nabla_\psi \log Z(\psi) = T(x) - \mathbf{G}_\psi[T].$$

By assumption (iv), $\nabla_\psi b(x, \psi) \in L^2(\mathbf{P})$, showing that condition (C10).

To show that the central limit theorem applies to $\hat{\psi}_{\mathrm{CE}}$, we additionally show that assumption (i) in Proposition 3.4 is fulfilled, which will finish the proof. To this end, note that

$$w(x) b'(x, \psi, \xi) = w(x) \xi^T \nabla_\psi b(x, \psi) = \xi^T \left(w(x)(T(x) - \mathbf{G}_\psi[T])\right) \in L^2(\mathbf{G})$$

by assumption (iii).

Finally, to show the representation of M, note that by Proposition 3.4 we have for $X \sim \mathbf{G}$

$$M = \operatorname{Cov}\left(w(X)\left(T(X) - G_{\psi_{\mathrm{CE}}}[T]\right)\right),$$

and $\mathbb{E} w(X)\left(T(X) - \mathbf{G}_{\psi_{\mathrm{CE}}}[T]\right) = 0$ as $\mathbf{G}_{\psi_{\mathrm{CE}}}[T] = \mathbf{P}[T]$. □

The form of the asymptotic covariance matrix is that of the sandwich estimator (White, 1982), corrected for the importance sampling with $\mathbf{G}$. This is not surprising: the CE-method essentially performs maximum likelihood estimation of ψ where the data comes from the misspecified $\mathbf{P}$. Additionally, we have to correct the variance for performing importance sampling with $\mathbf{G}$, instead of sampling directly from $\mathbf{P}$.

The assumptions of Theorem 3.6 are minimal to facilitate the proof. The existence and positive definiteness of the Fisher information matrix are easily checked for the exponential family proposal and hold for minimal regular exponential families. Additionally, we have two moment constraints that involve the weights w and the sufficient statistic T. That $wT \in L^2(\mathbf{G})$ may be seen as a generalization of the existence of the second moment $\rho = \mathbf{G}[w^2]$, adapted to the exponential family setting. As such it is a natural requirement. That $T \in L^2(\mathbf{P})$ is required for the application of Theorem 3.7, and, as mentioned before, should not be problematic in practice, except for heavy-tailed distributions.

For our application, we will choose $(\mathbf{G}_\psi)_{\psi \in \Psi}$ to consist of Gaussian distributions with natural parameter $\psi = \left(\Sigma^{-1}\mu, -\frac{1}{2}\Sigma^{-1}\right)$ and sufficient statistic $T(x) = \left(x, xx^T\right)$. Thus $T \in L^2(\mathbf{P})$ is equivalent to $\mathbf{P}$ having fourth order moments, which is reasonable if the target is not heavy-tailed.

If $(\mathbf{G}_\psi)_{\psi \in \Psi}$ do not form an exponential family, $\hat{\psi}_{\mathrm{CE}}$ will still be consistent and asymptotically normal, provided the usual regularity conditions for M-estimators apply. These usually include conditions to ensure the maximum is well-separated and the target is sufficiently smooth such that a Taylor expansion around the maximum is feasible. To extend our results to more involved settings, we refer the reader to (Van der Vaart, 2000) for an empirical process treatment of M- and related Z-estimators, (Haberman, 1989) for asymptotics when the objective function is concave, but the maximum may lie on the border of the parameter space and (Liang and Zeger, 1995) for a review of estimators based on estimating equations.

However, these conditions will become more intricate than the ones we have provided here, as the concavity of the log densities is a rather strong property. As a result, we expect that assessing whether these conditions are satisfied in practice be more difficult.

3.4.3 Efficient Importance Sampling (EIS)

Efficient Importance Sampling (EIS) (Richard and Zhang, 2007) provides an alternative to the CE-method. Instead of minimizing the KL-divergence between the target $\mathbf{P}$ and proposal $\mathbf{G}_\psi, \psi \in \Psi$, EIS aims at minimizing the variance of the logarithm of importance sampling weights. Our discussion of (Chatterjee and Diaconis, 2018), Theorem 3.2, especially Lemma 3.5, suggests that this is worthwhile. Thus, EIS finds ψ_{EIS} which is a feasible solution to the following optimization problem

$$\min_{\psi \in \Psi} \operatorname{Var}_{\mathbf{P}} [\log w_\psi] = \min_{\psi \in \Psi} \mathbf{P} \left[\log w_\psi - \mathbf{P} \log w_\psi\right]^2, \tag{3.17}$$

where, as in the last section, $\log w_\psi = \log p - \log g_\psi$.

Two problems arise: $\mathbf{P}[\log w_\psi] = \mathcal{D}_{\text{KL}}(\mathbf{P}||\mathbf{G}_\psi)$ is usually intractable and we usually only have access to the unnormalized weights $\frac{\tilde{w}_\psi}{\mathbf{G}_\psi[w_\psi]} = w_\psi$, with unknown integration constant $\mathbf{G}_\psi[w_\psi]$. Both can be dealt with by introducing the nuisance parameter $\lambda = \mathbf{P}[\log \tilde{w}_\psi]$, utilizing the fact that the mean is the minimizer of the squared distance functional with the minimum value equal to the variance, should it exist. Indeed

$$\log w_\psi - \mathbf{P}[\log w_\psi] = \log \tilde{w}_\psi - \log \mathbf{G}_\psi[\tilde{w}_\psi] - \mathbf{P}[\log \tilde{w}_\psi] + \log \mathbf{G}_\psi[\tilde{w}_\psi] = \log \tilde{w}_\psi - \mathbf{P}[\log \tilde{w}_\psi],$$

so

$$\min_{\psi \in \Psi} \mathbf{P} \left[\log w_\psi - \mathbf{P}[\log w_\psi]\right]^2 = \min_{\psi \in \Psi, \lambda \in \mathbf{R}} \mathbf{P} \left[\log \tilde{w}_\psi - \lambda\right]^2,$$

where $\psi \in \Psi$ is a minimizer of the left-hand side if, and only if, $(\psi, \lambda) \in \Psi \times \mathbf{R}$ with $\lambda = \mathbf{P}[\log \tilde{w}_\psi]$ is a minimizer of the right-hand side.

Similar to the CE-method we restrict our in-depth analysis to natural exponential family proposals where

$$\log g_\psi(x) = \psi^T T(x) - \log Z(\psi).$$

In this case the optimization problem is reduced to

$$\min_{\psi \in \Psi, \lambda \in \mathbf{R}} \mathbf{P} \left[\log p - \psi^T T - \lambda\right]^2, \tag{3.18}$$

a weighted linear least squares problem. As we consider unnormalized weights $\tilde{w}$, we are additionally able to get rid of the potentially non-linear term $\log Z(\psi)$. Notice too that this is a convex objective function in ψ which, similar to the CE-method, will be very useful to derive asymptotics later on. For now, we begin with studying the existence and uniqueness of ψ_{EIS} similar to Proposition 3.2.

Lemma 3.9 (EIS for exponential families)**.** *Let* $(\mathbf{G}_\psi)_{\psi \in \Psi}$ *form a k-dimensional natural exponential family with log-densities*

$$\log g_\psi(x) = \psi^T T(x) - \log Z(\psi)$$

for $\Psi \subseteq \mathbf{R}^k$*. Suppose that* $\log p, T \in L^2(\mathbf{P})$*.*

If there is a $\psi_{EIS} \in \Psi$ *with*

$$\operatorname{Cov}_{\mathbf{P}}(T)\, \psi_{EIS} = \operatorname{Cov}_{\mathbf{P}}(T, \log p) \tag{3.19}$$

it is a global minimizer of Equation (3.17). If $\operatorname{Cov}_{\mathbf{P}}(T)$ *is non-singular,*

$$\psi_{EIS} = \operatorname{Cov}_{\mathbf{P}}(T)^{-1} \operatorname{Cov}_{\mathbf{P}}(T, \log p)$$

is the unique global minimizer.

Proof. Under the proposed conditions, we may consider Equation (3.18) instead, where the moment conditions on $\log p$ and T ensure that the problem is well-posed, i.e. the target is finite for all $\psi \in \Psi$. Thus the optimal $(\psi_{\text{EIS}}, \lambda_{\text{EIS}})$ are given by the best linear unbiased predictor (BLUP) of $\log p$ by the sufficient statistic T under $\mathbf{P}$ for ψ_{EIS} and $\mathbf{P}[\log \tilde{w}_{\psi_{\text{EIS}}}]$ for λ_{EIS}. The BLUP is given by any solution of

$$\operatorname{Cov}_{\mathbf{P}}(T)\psi_{\text{EIS}} = \operatorname{Cov}_{\mathbf{P}}(T, \log p),$$

cite something

i.e. ψ_{EIS} as stated in the lemma. Furthermore, if $\text{Cov}_{\mathbf{P}}(T)$ is non-singular, the solution to this equation is unique.

□

As the optimal ψ_{EIS} depends on several unknown quantities, EIS proceeds like the CE-method and employs importance sampling with a proposal $\mathbf{G}$, estimating ψ_{EIS} by

$$\left(\hat{\lambda}, \hat{\psi}_{\text{EIS}}\right) = \text{argmin}_{\lambda,\psi} \hat{\mathbf{P}}_N \left[\log \tilde{w}_\psi - \lambda\right]$$

where $X^1, \ldots, X^N \overset{\text{i.i.d}}{\sim} \mathbf{G}$. Again, if $\mathbf{G}_\psi, \psi \in \Psi$ form an exponential family with natural parameter ψ, this optimization problem turns into a weighted least squares problem, so we can estimate ψ_{EIS} with the standard weighted least squares estimator

$$\left(\hat{\lambda}', \hat{\psi}_{\text{EIS}}\right) = \left(\mathbf{X}^T \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{W} y$$

where the random design matrix $\mathbf{X}$[5] and diagonal weights matrix $\mathbf{W}$ are given by

$$\mathbf{X} = \begin{pmatrix} 1 & T(X^1)^T \\ \ldots & \ldots \\ 1 & T(X^N)^T \end{pmatrix}$$

and

$$\mathbf{W} = \text{diag}\left(W^1, \ldots, W^N\right),$$

and the observations are

$$y = \left(\log p(X^1), \ldots, \log p(X^N)\right)^T \in \mathbf{R}^N.$$

Alternatively, replacing $\mathbf{P}$ by $\hat{\mathbf{P}}_N$ in Equation (3.19), we obtain the equivalent formulation

$$\hat{\psi}_{\text{EIS}} = \text{Cov}_{\hat{\mathbf{P}}_N}(T)^{-1} \text{Cov}_{\hat{\mathbf{P}}_N}(T, \log p), \tag{3.20}$$

as long as $\text{Cov}_{\hat{\mathbf{P}}_N} T$ is non-singular.

An attractive feature of EIS is that if the target $\mathbf{P}$ is a member of the exponential family of proposals, i.e. there is a $\psi_{\mathbf{P}} \in \Psi$ such that $\mathbf{P} = \mathbf{G}_{\psi_{\mathbf{P}}}$, then EIS finds the optimal $\psi_{\text{EIS}} = \psi_{\mathbf{P}}$ a.s. for a finite number of samples.

Proposition 3.5 (Finite sample convergence of EIS)**.** *Suppose $\mathbf{G}_\psi, \psi \in \Psi \subseteq \mathbf{R}^k$ for a natural exponential family w.r.t. Lebesgue measure, where the support of the sufficient statistic* $\text{supp}\, T$ *is open in $\mathbf{R}^k$. Furthermore let $\mathbf{G}$ be a probability measure on $\mathbf{R}^m$ that is equivalent to $\mathbf{P}$, i.e. $\mathbf{G} \ll \mathbf{P}$ and $\mathbf{P} \ll \mathbf{G}$.*

If there is a $\psi_{\mathbf{P}} \in \Psi$ such that $\mathbf{P} = \mathbf{G}_{\psi_{\mathbf{P}}}$, then $\hat{\psi}_{EIS} = \psi_{\mathbf{P}}$ a.s. for $N \geq k$.

Proof. As $\mathbf{P}$ stems from the same exponential family as $\mathbf{G}_\psi$, the pseudo-observations are

$$\log p = \psi_{\mathbf{P}}^T T - \log Z(\psi_{\mathbf{P}}).$$

Thus $\text{Cov}_{\hat{\mathbf{P}}_N}(T, \log p) = \text{Cov}_{\hat{\mathbf{P}}_N}(T)\, \psi_{\mathbf{P}}$. If we can show that $\text{Cov}_{\hat{\mathbf{P}}_N} T$ is non-singular, Equation (3.20) implies that $\hat{\psi}_{\text{EIS}} = \psi_{\mathbf{P}}$ a.s..

[5] if $\mathbf{XWX}$ is not invertible, replace the inverse by the Moore-Penrose pseudoinverse

If $\mathrm{Cov}_{\hat{\mathbf{P}}_N} T$ were singular, there would exist a $\psi \in \mathbf{R}^k$ such that

$$\psi^T \mathrm{Cov}_{\hat{\mathbf{P}}_N}(T)\psi = \mathrm{Cov}_{\hat{\mathbf{P}}_N}\left(\psi^T T\right) = 0.$$

In this case the a.s. non-zero $W^i(X^i)T(X^i)$ would lie in the orthogonal complement $\psi^\perp$ for all $i = 1, \dots, N$. As the weights are a.s. positive by the assumed equivalence of $\mathbf{G}$ and $\mathbf{P}$, the same holds true for $T(X^i), i = 1, \dots, N$. If N is bigger than k, the probability that this happens is 0, as $\operatorname{supp} T$ is open. Thus $\mathrm{Cov}_{\hat{\mathbf{P}}_N} T$ is non-singular almost surely and the result is shown. □

Note that if in the above proposition only $\mathbf{G}_\psi \gg \mathbf{P}$ holds, we obtain, by a similar argument, that

$$\mathbb{P}\left(\hat{\psi}_{\mathrm{EIS}} = \psi_{\mathbf{P}}\right) \overset{N\to\infty}{\longrightarrow} 1.$$

Additionally, we then have to take care of the event $\{w(X) = 0\}$, whose probability is now potentially positive.

We now turn to deriving asymptotics for $\hat{\psi}_{\mathrm{EIS}}$. As for the CE-method, we start with proving that $\hat{\psi}_{\mathrm{EIS}}$ consistently estimates ψ_{EIS}. For this we need to ensure that ψ_{EIS} is the unique solution to Equation (3.17), as otherwise, consistent estimators of ψ_{EIS} cannot exist. As Equation (3.18) is a linear least squares problem, the objective function is convex, and so we can apply Theorem 3.5 and Proposition 3.3.

Theorem 3.8 (consistency of $\hat{\psi}_{\mathrm{EIS}}$)**.** *Let $(\mathbf{G}_\psi)_{\psi\in\Psi}$ form a k-dimensional natural exponential family with log-densities*

$$\log g_\psi(x) = \psi^T T(x) - \log Z(\psi)$$

for convex $\Psi \subseteq \mathbf{R}^k$. Let $\mathbf{G} \gg \mathbf{P}$ be a proposal and suppose that

(i) $\log p, T \in L^2(\mathbf{P})$ *and*

(ii) $\mathrm{Cov}_{\mathbf{P}}(T)$ *is non-singular,*

(iii) $\psi_{EIS} \in \operatorname{int}\Psi$.

Then

$$\hat{\psi}_{EIS} \overset{N\to\infty}{\longrightarrow} \psi_{EIS}$$

almost surely.

Proof. We follow the same strategy as in the proof of Theorem 3.4. Let

$$b : \mathbf{R}^p \times \mathbf{R}^{k+1} \to [-\infty, \infty) \qquad b(x, \psi') = \begin{cases} -\frac{1}{2}\left(\log p(x) - \psi^T T(x) - \lambda\right)^2 & \psi \in \Psi \\ -\infty & \text{else,} \end{cases}$$

where $\psi' = (\psi, \lambda) \in \mathbf{R}^{k+1}$. For fixed x this function is concave, as its Hessian is negative semi-definite:

$$H_{\psi'} b(x, \psi') = -\begin{pmatrix} 1 & T(x)^T \\ T(x) & T(x)T(x)^T \end{pmatrix} = -\begin{pmatrix} 1 & T(x)^T \end{pmatrix}\begin{pmatrix} 1 & T(x)^T \end{pmatrix}^T,$$

if $\psi \in \Psi$. Let $X^1, \dots, X^N \overset{\text{i.i.d}}{\sim} \mathbf{P}$ and let $\tilde{\mathbf{P}}_N$ be their empirical distribution. For $\psi \in \Psi, \lambda \in \mathbf{R}$ we have

$$\mathbf{P}\left[b(\cdot, \psi')\right] = -\frac{1}{2}\mathbf{P}\left[\left(\log p - \psi^T T - \lambda\right)^2\right] < \infty,$$

as $\log p, T \in L^2(\mathbf{P})$. Let us now check that conditions (C1) - (C3) are fulfilled.

(C1) is fulfilled, as we assumed $\psi_{\mathrm{EIS}} \in \operatorname{int}\Psi$. (C2) holds, as ψ_{EIS} is the unique global maximizer by Lemma 3.9, as $\mathrm{Cov}(T)$ is non-singular. (C3) obviously holds.

Thus $\hat{\psi}_{\mathrm{EIS}}$ is strongly consistent if $\mathbf{G} = \mathbf{P}$. If $\mathbf{G}$ is different from $\mathbf{P}$, we can apply Proposition 3.3, where the existence of M-estimators is ensured by Equation (3.20), using the Moore-Penrose inverse if $\mathrm{Cov}_{\hat{\mathbf{P}}_N}(T)$ is singular. □

discussion of assumptions

As Equation (3.20) expresses $\hat{\psi}_{\text{EIS}}$ in terms of empirical covariances, we could alternatively prove consistency by ensuring that the empirical covariances are consistent as well, for which we would need to ensure that fourth-order moments of $\log p$ and T w.r.t. $\mathbf{P}$ exist. This strategy may be fruitful if ψ_{EIS} does not lie in the interior of Ψ, although the more sophisticated treatment of (Haberman, 1989) may also be applicable under these circumstances.

Additionally, if fourth-order moments exist, we can derive a central limit theorem, similar to Theorem 3.6, for EIS.

Theorem 3.9 (CLT for $\hat{\psi}_{\text{EIS}}$)**.** *Let $(\mathbf{G}_\psi)_{\psi\in\Psi}$ form a k-dimensional natural exponential family with log-densities*

$$\log g_\psi(x) = \psi^T T(x) - \log Z(\psi),$$

and convex parameter space $\Psi \subseteq \mathbf{R}^k$. Let $\mathbf{G} \gg \mathbf{P}$ be a proposal with weights $w = \frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{G}}$.

(i) $wT_iT_j, w(\log p)^2 \in L^2(\mathbf{G})$ *for* $i = 1,\dots,k$, $j = 1,\dots,k$,

(ii) $\log p, T_i \in L^4(\mathbf{P})$ *for all* $i = 1,\dots,k$

(iii) $\operatorname{Cov}_{\mathbf{P}}(T)$ *is non-singular and* $\psi_{EIS} \in \operatorname{int}\Psi$.

Then

$$\sqrt{N}(\hat{\psi}_{EIS} - \psi_{EIS}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, BMB)$$

where $B = \operatorname{Cov}_{\mathbf{P}}(T)^{-1}$ *and*

$$M = \operatorname{Cov}_{\mathbf{G}}\left(w\left(\log p - {\psi_{EIS}}^T T - \lambda_{EIS} - \mathbf{P}[T]\right)T\right).$$

Proof. Similar to the proof of Theorem 3.6, we combine Theorem 3.7 and Proposition 3.4. Let

$$b : \mathcal{X} \times \mathbf{R}^{k+1} \to [-\infty, \infty) \qquad b(x,\psi') \begin{cases} -\frac{1}{2}\left(\log p(x) - {\psi'}^T T'(x)\right) & x \in \Psi \\ -\infty & \text{else,} \end{cases}$$

where $\psi' = (\psi, \lambda) \in \Psi \times \mathbf{R}$ and $T'(x) = \begin{pmatrix} T(x) & 1 \end{pmatrix}$. For $\psi \in \Psi$ the map $(\psi,\lambda) \to \mathbf{P}[b(\cdot,(\psi,\lambda))]$ is differentiable with gradient

$$\nabla_{\psi'}\mathbf{P}[b(\cdot,\psi')] = -\mathbf{P}\left[\left(\log p - {\psi'}^T T'\right)T'\right] = \begin{pmatrix} -\mathbf{P}\left[T'\log p - T'T'^T\psi'\right] \\ -\mathbf{P}\left[\log p - {\psi'}^T T'\right] \end{pmatrix}$$

and Hessian

$$H_{\psi'}\mathbf{P}[b(\cdot,\psi')] = -\mathbf{P}\left[T'T'^T\right] = -\begin{pmatrix} \mathbf{P}\left[TT^T\right] & \mathbf{P}[T^T] \\ \mathbf{P}[T] & 1 \end{pmatrix}.$$

The Hessian is negative definite, as for all $\psi \in \mathbf{R}^k, \lambda \in \mathbf{R}$ we have

$$\begin{aligned} \begin{pmatrix} \psi^T & \lambda \end{pmatrix} H_{\psi'}\mathbf{P}\left[b(\cdot,\psi')\right]\begin{pmatrix} \psi^T & \lambda \end{pmatrix}^T &= -\left(\psi^T \operatorname{Cov}_{\mathbf{P}}(T)\psi + \psi^T\mathbf{P}[T]\mathbf{P}[T]^T\psi + 2\psi^T\mathbf{P}[T]\lambda + \lambda^2\right) \\ &= -\left(\psi^T \operatorname{Cov}_{\mathbf{P}}(T)\psi + (\lambda + \psi^T\mathbf{P}[T])^2\right) \le 0, \end{aligned}$$

with equality if, and only if, both λ and ψ are 0, as $\operatorname{Cov}_{\mathbf{P}}(T)$ is assumed to be positive definite. Thus condition (C7) is fulfilled.

For condition (C10), we can verify that for all $i, j = 1,\dots,k+1$

$$(\nabla_{\psi'}b(\cdot,\psi'))_i(\nabla_{\psi'}b(\cdot,\psi'))_j = \left(\log p - {\psi'}^T T'\right)^2 T'_iT'_j$$

is in $L^1(\mathbf{P})$ by assumption (ii) and the Hölder inequality.

To apply Proposition 3.4 we need to show that $w(\cdot)b'(\cdot,\psi',\xi') \in L^2(\mathbf{G})$ for all $\xi' \in \mathbf{R}^{k+1}$ and all ψ' in a neighborhood of ψ_{EIS}, for this it suffices that we show

$$w^2(\nabla_{\psi'}b(\cdot,\psi'))_i(\nabla_{\psi'}b(\cdot,\psi'))_j = w^2\left(\log p - {\psi'}^T T'\right)^2 T'_iT'_j$$

is in $L^1(\mathbf{G})$, which holds, again, by assumption Item (i) and the Hölder inequality.

We have thus shown a central limit theorem for $\hat{\psi}'_{\text{EIS}} = \left(\hat{\psi}_{\text{EIS}}, \hat{\lambda}_{\text{EIS}}\right)$, i.e.

$$\sqrt{N}\left(\hat{\psi}'_{\text{EIS}} - \psi_{\text{EIS}}\right) \to \mathcal{N}(0, M'B'M')$$

with $B' = -\left(H_{\psi'_{\text{EIS}}}\mathbf{P}\left[b(\cdot, \psi'_{\text{EIS}})\right]\right)^{-1}$ and $M' = \text{Cov}\left(w(X)\nabla_{\psi'_{\text{EIS}}} b(X, \psi'_{\text{EIS}})\right)$ for $X \sim \mathbf{G}$. By using the inversion formula for block matrices, we obtain

$$\begin{aligned} B' &= \begin{pmatrix} \mathbf{P}\left[TT^T\right] & \mathbf{P}[T^T] \\ \mathbf{P}[T] & 1 \end{pmatrix}^{-1} = \begin{pmatrix} \Sigma + \mu\mu^T & \mu^T \\ \mu & 1 \end{pmatrix}^{-1} \\ &= \begin{pmatrix} (\Sigma + \mu\mu^T - \mu\mu^T)^{-1} & 0 \\ 0 & 1 - \mu^T(\Sigma + \mu\mu^T)^{-1}\mu \end{pmatrix} \begin{pmatrix} I_k & -\mu^T \\ -\mu(\Sigma + \mu\mu^T)^{-1} & 1 \end{pmatrix} \\ &= \begin{pmatrix} \Sigma^{-1} & -\mu^T\Sigma^{-1} \\ -\Sigma^{-1}\mu & 1 - \mu^T\left(\Sigma + \mu\mu^T\right)^{-1}\mu \end{pmatrix} \end{aligned}$$

where $\Sigma = \text{Cov}_{\mathbf{P}}(T)$ and $\mu = \mathbf{P}[T]$. Similarly,

$$M' = \begin{pmatrix} \text{Cov}_{\mathbf{G}}\left(wW_{\psi_{\text{EIS}}}T\right) & \text{Cov}_{\mathbf{G}}\left(wW_{\psi_{\text{EIS}}}T, wW_{\psi_{\text{EIS}}}\right) \\ \text{Cov}_{\mathbf{G}}\left(wW_{\psi_{\text{EIS}}}, wW_{\psi_{\text{EIS}}}T\right) & \text{Cov}_{\mathbf{G}}\left(wW_{\psi_{\text{EIS}}}\right) \end{pmatrix},$$

where $W_{\psi_{\text{EIS}}} = \log p - \psi'^T_{\text{EIS}}T'$.

If $\mu \neq 0$, we may change the sufficient statistic of the exponential family such that this holds, i.e. let $\tilde{T} = T - \mathbf{P}[T]$, then

$$\log g_\psi(x) = \psi^T T(x) - \log Z(\psi) = \psi^T\tilde{T}(x) - \log\tilde{Z}(\psi)$$

where $\tilde{Z}(\psi) = \log Z(\psi) + \mathbf{P}[T]$. As ψ_{EIS}, Equation (3.19), only depends on $T - \mathbf{P}[T]$ under $\mathbf{P}$, this does not change ψ_{EIS}. Similarly, $\hat{\psi}_{\text{EIS}}$, Equation (3.20), is unaffected by subtracting a constant from T. Only

$$\tilde{\lambda}_{\text{EIS}} = \lambda_{\text{EIS}} + \mathbf{P}[T]$$

and similarly $\hat{\lambda}_{\text{EIS}}$ are changed.

Thus, without loss of generality, we may assume that $\mathbf{P}[T] = 0$. Then

$$B' = \begin{pmatrix} \Sigma^{-1} & 0 \\ 0 & 1 \end{pmatrix}$$

is a diagonal matrix. Taking the ψ_{EIS} marginal of the asymptotic normal distribution, we arrive at

$$\sqrt{N}\left(\hat{\psi}_{\text{EIS}} - \psi_{\text{EIS}}\right) \to \mathcal{N}\left(0, BMB\right)$$

with $B = \text{Cov}_{\mathbf{P}}(T)$ and $M = \text{Cov}_G\left(w\left(\log p - {\psi_{\text{EIS}}}^T T - \lambda_{\text{EIS}} - \mathbf{P}[T]\right)T\right)$, as promised.

□

3.5 Interim discussion

Before we apply EIS and the CE-method in the SSM context, let us consolidate what we have achieved by the asymptotic analysis in the preceding two subsections and reason which of the two methods should be used in which circumstances.

We start with a discussion of the optimal values ψ_{CE} and ψ_{EIS}. Notice that ψ_{EIS} depends on second-order moments of the sufficient statistic T, as well as the shape of $\log p$, whereas the optimal parameter for the CE-method ψ_{CE} depends only on the first-order moments of T. This dependence

on higher-order moments may be beneficial for the EIS method, for example, if the covariance of T under $\mathbf{P}$ is very different from that under $\mathbf{G}_\psi$.

should have an example for this later

The two methods differ concerning the assumptions that are required for uniqueness, consistency and the central limit theorem to hold if the proposals come from an exponential family. For uniqueness, Proposition 3.2 and lemma 3.9, both methods require that the covariance of T is non-singular, however, the measures under which the covariance are considered differ: for the CE-method we need $\mathrm{Cov}_{\mathbf{G}_{\psi_{\mathrm{CE}}}}(T)$ to be non-singular, while for EIS the same has to hold for $\mathrm{Cov}_{\mathbf{P}}(T)$. While the former is easy to ensure, the latter depends on the intractable target $\mathbf{P}$ and may be more difficult to verify in practice, depending on T.

Regarding the consistency results, Theorems 3.4 and 3.8 as well as the central limit theorems, Theorems 3.6 and 3.9, EIS requires that the sufficient statistic be twice as often $\mathbf{P}$-integrable as the CE-method. Additionally, the EIS results assume that $\log p$ is sufficiently often $\mathbf{P}$-integrable. Therefore, EIS is, at first glance, more restrictive than the CE-method. However, our application will perform importance sampling with Gaussian proposals where $T(x) = \begin{pmatrix} x \\ xx^T \end{pmatrix}$. For importance sampling to be consistent in this setting, we assume that the target has thinner tails than the Gaussian proposal, which implies that all polynomial moments of the target, and thus of T exist. A similar argument can be made for $\log p$, and so the assumptions are likely to be fulfilled when Gaussian importance sampling is consistent.

To compare the asymptotic covariance matrices of both methods, note that both covariance matrices have the same „bread-meat-bread“ factorization, as they are asymptotic covariance matrices of M-estimators. We see that both $B_{\mathrm{CE}} = I(\psi) = \mathrm{Cov}_{G_{\psi_{\mathrm{CE}}}}(T)^{-1}$ and $B_{\mathrm{EIS}} = \mathrm{Cov}_{\mathbf{P}}(T)^{-1}$ are precision matrices of the sufficient statistic T, one with respect to the optimal CE-method proposal and one with respect to the target. Thus, if $\mathbf{P}$ is well approximated by $\mathbf{G}_{\psi_{\mathrm{CE}}}$, we would expect these two components to be close to one another. For $M_{\mathrm{CE}} = \mathrm{Cov}_{\mathbf{G}}(wT)$ and $M_{\mathrm{EIS}} = \mathrm{Cov}_{\mathbf{G}}\left(w(\log p - \psi_{\mathrm{EIS}}^T T - \lambda_{\mathrm{EIS}} - \mathbf{P}[T])T\right)$, there is a more notable difference, i.e. the presence of the $\log p - \psi_{\mathrm{EIS}}^T T - \lambda_{\mathrm{EIS}}$ term. If the EIS approximation performs well, we can expect this term to be small, as it is the prediction error of the least squares approximation of $\log g_\psi$ to $\log p$. Therefore, we expect that EIS outperforms the CE-method in terms of asymptotic variance in these settings. In agreement with Proposition 3.5, $M_{\mathrm{EIS}} = 0$ if $\log p = \log g_{\psi_{\mathbf{P}}}$ so that $\psi_{\mathrm{EIS}} = \psi_{\mathbf{P}}$.

Additionally, both M_{CE} and M_{EIS} depend on the proposal $\mathbf{G}$, and indicate how one might tailor the initial proposal $\mathbf{G}$ to produce low-variance estimates. For the CE-method we might choose $\mathbf{G}$ such that the trace determinant of $\mathbf{G}[w^2 TT^T]$ becomes small. This is not necessarily achieved by the CE-method proposal $\mathbf{G}_{\hat{\psi}_{\mathrm{CE}}}$, and so it may be worthwhile to investigate using two types of proposals in the CE-method, one that makes M_{ce} small and $\mathbf{G}_{\psi_{\mathrm{CE}}}$. This is especially relevant as our simulation studies, Section 3.8, suggest that the asymptotic covariance of the CE-method is usually inferior to that of EIS. For EIS, a similar approach might be fruitful, but is not as urgent as that for the CE-method, as the asymptotic covariance of EIS is usually small enough to be feasible in practice.

Finally, let us stress that these asymptotic considerations are, to the author's knowledge, novel results and should be straightforward to extend if the proposals $(\mathbf{G}_\psi)_{\psi\in\Psi}$ do not form a natural exponential family. As any minimal exponential family may be reduced to a natural exponential family by reparametrization, see (Brown, 1986, Theorem 1.9), the delta method can be used to derive CLTs in this case as well, as Proposition 3.2 and lemma 3.9 still apply. If the family is not minimal the optimal values ψ_{EIS} and ψ_{CE} may be non-unique, so we cannot hope to estimate them consistently. In this case the user should choose a minimal parametrization, see again (Brown, 1986, Theorem 1.9). For non-exponential family proposals our results should also carry over, provided the usual regularity conditions ensuring uniqueness, consistency and asymptotic normality for M-estimators hold. If the objective functions are not concave as they are in our setting one usually requires uniformly bounded third-order derivatives of the objective function to exist.

Furthermore, our results can also be extended to the so-called Variance-Minimization method

(VM-method) which determines an optimal proposal by solving the following optimization problem:

$$\min_{\psi\in\Psi} \mathrm{Var}_{\mathbf{G}_\psi}(w_\psi) = \min_{\psi\in\Psi} \mathbf{G}_\psi\left[w_\psi^2\right] = \min_{\psi\in\Psi} \mathbf{P}\left[w_\psi\right],$$

where the first equality holds as $\mathbf{G}_\psi[w_\psi] = 1$ for all ψ. Thus the VM-method chooses ψ such that the second moment of importance sampling weights, ρ, becomes small. Again, this is sensible by the discussion surrounding ρ and the ESS. Again, one uses importance sampling with a proposal $\mathbf{G}$ to approximate $\mathbf{P}[w_\psi]$ by $\hat{\mathbf{P}}_N[w_\psi]$, and solves this noisy version of the problem. Unfortunately, there is no closed form for the optimal ψ_{VM} or $\hat{\psi}_{\mathrm{VM}}$, even if the proposals form a natural exponential family. Still, as $x \mapsto w_\psi(x)$ is convex, so is $x \mapsto \mathbf{P}[w_\psi]$, and we can apply Theorems 3.5 and 3.7 in combination with Propositions 3.3 and 3.4 to show, under suitable regularity conditions, the consistency and asymptotic normality of the method.

Now that we have gained theoretical insight into optimal importance sampling, let us apply these insights to the SSMs that we are interested in.

3.6 Gaussian importance sampling for state space models

For the types of models considered in this thesis, importance sampling is used to infer the posterior distribution. Given a state space model of the form (3.1) and observations $Y = Y_{:n}$, let $\mathbf{P}$ be the distribution of the states $X = X_{:n}$, conditional on Y and f be a function of interest. The task at hand is now to find a suitable proposal $\mathbf{G}$, using the methods presented in the last section. If n is large, the posterior distribution lives in a high dimensional state of dimension $m \cdot n$, so to obtain $\mathbf{G}$ efficiently, we should exploit the available structure. Additionally, we want $\mathbf{G}$ to be tractable, so simulating from it is possible and evaluating the weights w up to a constant is possible.

The multivariate Gaussian distribution is a good candidate in this setting, as simulating from it is straightforward and its density can be evaluated analytically. However, naively performing the optimal importance sampling methods from the previous section for all multivariate Gaussians is computationally inefficient as the family of distributions has $\mathcal{O}((n \cdot m)^2)$ many parameters. We can, however, exploit the available structure of the SSM to find parameterizations with fewer parameters by either using smoothing distributions of GLSSMs (Section 3.6.1) or approximating with a Gaussian discrete-time Markov process (Section 3.6.2).

Using Gaussian proposals, while computationally efficient, also comes with some drawbacks. The whole procedure hinges on the assumption that there is a Gaussian that is close to the target distribution. In the setting of SSMs this is not guaranteed, as the targets may contain multiple modes or heavy tails, features that may, in the worst case, lead to inconsistent importance sampling estimates. Additionally, even if there is a Gaussian distribution that facilitates consistent importance sampling, finding it in practice may be complicated, as the proposals generated by the LA, CE-method and EIS have deteriorating performance for fixed sample size N (in terms of ESS and convergence) with increasing dimension, see Section 3.8.5.

small lit. review

3.6.1 The GLSSM-approach

The first approach is motivated by the fact that the target posterior is again a Markov process, as are posteriors in GLSSMs. Additionally, the posterior distribution in GLSSMs is again Gaussian, and straightforward to simulate from by, e.g., the FFBS algorithm (Algorithm 3) or the simulation smoother (Durbin and Koopman, 2002). Thus parameterizing the proposals $\mathbf{G}$ by the posterior of a suitably chosen GLSSM may be a fruitful approach. For the models we consider in this thesis, the distribution of states is already Gaussian and the observations are conditionally independent given the states. Thus a natural GLSSM to use as a proposal consists of keeping the prior distribution of states and replacing the distribution of observations with conditionally independent Gaussian distributions and the actual observations by synthetic ones. By the assumed conditional independence, this model only needs $2p \cdot (n+1)$ many parameters, $p \cdot (n+1)$ for the synthetic

observations and $p \cdot (n+1)$ for their variances. We term this approach the **GLSSM-approach** to importance sampling.

In total, the GLSSM-approach considers parametric proposals $\mathbf{G}_\psi$ of the form

$$\begin{aligned} \mathbf{G}_\psi &= \mathcal{L}(X|Z=z), \\ Z_t &= B_t X_t + \eta_t, \\ \eta_t &\sim \mathcal{N}(0, \Omega_t), \\ \Omega_t &= \operatorname{diag}\left(\omega_t^2\right) = \operatorname{diag}\left(\omega_{t,1}^2, \dots, \omega_{1,p}^2\right). \end{aligned} \tag{3.21}$$

where the distribution of X is given by (3.4), $\psi = \left(z, \omega^2\right)$ for $z = (z_0, \dots, z_n) \in \mathbf{R}^{n \times m}$ and $\omega^2 = \left(\omega_0^2, \dots, \omega_n^2\right) \in \mathbf{R}^{n \times m}$. Alternatively the natural parametrization $\psi = \left(z \oslash \omega^2, -1 \oslash \left(2\omega^2\right)\right)$ may also be used, where $\oslash$ is the Hadamard, i.e. entry-wise, division. Simulation from $\mathbf{G}_\psi$ may be efficiently implemented by the FFBS algorithm, as $\mathbf{G}_\psi$ is the smoothing distribution of a GLSSM.

In this setting, the importance sampling weights are given by

$$w(x) = \frac{p(x|y)}{g(x|z)} = \frac{p(y|x)p(x)}{g(z|x)p(x)} \frac{g(z)}{p(y)} \propto \prod_{t=0}^{n} \frac{p(y_t|x_t)}{g(z_t|x_t)},$$

so they can be computed efficiently. Additionally, for a LCSSM with linear signals, $p(y_t|x_t)$ and $g(z_t|x_t)$ depend on x_t only through the signal $s_t = B_t x_t$, and we have

$$w(x) \propto \prod_{t=0}^{n} \frac{p(y_t|s_t)}{g(z_t|s_t)}, \tag{3.22}$$

which implies that auto-normalized weights may be calculated by using the signal smoother (Jungbacker and Koopman, 2007, Theorem 2). As (Durbin and Koopman, 2012) (Durbin and Koopman, 2012, Section 4.5.3) argue, it is often computationally more efficient to treat only on the signals $S_{:n}$ instead of the states $X_{:n}$, the idea being that the dimension of S_t, p, is usually much smaller than that of X_t, m.

As the joint distribution of (X, S) is a Gaussian distribution, by Lemma 3.1 $X|S = s$ is again Gaussian, with known conditional mean and covariance matrix and density $p(x|s) = g(x|s)$. If $(\tilde{X}_t)_{t=0,\dots,n}$ is a draw from this conditional distribution a quick calculation reveals that a.s. $B_t \tilde{X}_t = S_t$, and so, as expected, the weights $w(\tilde{X}_t)$ are a.s. constant and given by (up to the integration constant) Equation (3.22). Producing a draw from this conditional distribution can be achieved by the FFBS algorithm (Algorithm 3), as (X, S) form a GLSSM with degenerate observation covariance matrices $\Omega_t = 0$.

By the assumed conditional independence of observations given signals, we have

$$p(x, s|y) \propto p(x|s)p(s|y),$$

and so if one is interested in the states, rather than the signals, importance sampling with the proposal Equation (3.21) can be achieved in a two-step procedure: first sample from $g(s|z)$, then run the FFBS algorithm to sample from $g(x|s) = p(x|s)$ using the same weights for MC-integration.

The GLSSM-approach is the standard approach for finding the LA in LCSSM (Durbin and Koopman, 2012; Durbin and Koopman, 1997) and may even be applied when the observation densities are not log-concave(Jungbacker and Koopman, 2007). The approach also leads to efficient implementation for EIS (Koopman, Lit, and Nguyen, 2019). However, as will become apparent in the later part of this section, it is infeasible for the CE-method if n is large.

We now give a concise overview over how to perform the LA and EIS for LCSSM, but refer the reader for more details to the respective literature. The LA

elaborate on LA / EIS, Durbin Koopman book, EIS, MEIS, NEIS papers

Algorithm 5 The LA for LCSSM

Algorithm 6 EIS for LCSSM

For the CE-method, using the GLSSM-approach turns out to be difficult numerically. For a high-level argument of why this is true, let us ignore the Markov structure of the model for the moment. As the CE-method matches moments of the target and proposal, applying it to fit model (3.21) amounts to matching the moments of $\mathbf{G}_\psi$ to those of the target posterior $\mathcal{L}(X|Y=y)$ in the SSM. Unfortunately, the covariance of $\mathbf{G}_\psi$ is given by $\left(\Sigma^{-1}+B^T\Omega^{-1}B\right)^{-1}$, where Σ is the covariance of all states, $B=\text{block-diag}(B_0,\ldots,B_n)$ and $\Omega=\text{block-diag}\left(\Omega_0,\ldots,\Omega_n\right)$. Choosing the diagonal matrix Ω such that the covariance of $\mathbf{G}_\psi$ matches this expression is numerically expensive: we either need to invert the large (dimension $(n+1)m\times(n+1)m$) covariance matrix, or solve numerically for the $(n+1)p$ parameters. The problem at hand is that we cannot decouple this into $(n+1)$ equations of dimension p as we did for EIS, because all entries of $(\Sigma^{-1}+B^T\Omega^{-1}B)^{-1}$ depend on all entries of Ω.

To make matters more concrete, the CE-method finds $\psi=(z,\omega^2)$ such that model (3.21) maximizes the cross entropy with the target $\mathbf{P}^{X|Y=y}$. For simplicity, let us assume that $m=p$, B is the identity and we only observe a single y. Using Lemma 3.1, we see that when $X\sim\mathcal{N}(\mu,\Sigma)$, the conditional distribution of X given $Z=z$, $\mathbf{G}_\psi$, is a Gaussian distribution with mean $\tilde{\mu}=\mu+\Sigma\left(\Sigma+\Omega\right)^{-1}(z-\mu)$ and covariance matrix $\tilde{\Sigma}=\left(\Sigma^{-1}+\Omega^{-1}\right)^{-1}$ for $\Omega=\text{diag}\left(\omega^2\right)$, where $\omega^2>0$. Assuming that Σ is non-singular, we can reparameterize the objective function of the CE-method by $\tilde{\mu}$,

$$\begin{aligned}\max_{z,\omega^2}\int p(x|y)\log g_\psi(x|z)\mathrm{d}x &= \max_{\tilde{\mu},\omega^2}\int p(x|y)\left(-\frac{1}{2}(x-\tilde{\mu})^T\tilde{\Sigma}^{-1}(x-\tilde{\mu})-\frac{1}{2}\log\det\tilde{\Sigma}\right)\mathrm{d}x\\ &= \max_{\tilde{\mu},\omega^2}-\frac{1}{2}(\gamma-\tilde{\mu})^T\tilde{\Sigma}^{-1}(\gamma-\tilde{\mu})-\frac{1}{2}\text{trace}\left(\tilde{\Sigma}^{-1}\Gamma\right)-\frac{1}{2}\log\det\tilde{\Sigma},\end{aligned} \tag{3.23}$$

where $\gamma=\mathbb{E}\left(X|Y=y\right)$ and $\Gamma=\text{Cov}\left(X|Y=y\right)$. Thus the optimal $\tilde{\mu}$ is γ and to find the optimal ω^2 we have to minimize

$$\text{trace}\left(\left(\Sigma^{-1}+\Omega^{-1}\right)\Gamma\right)-\log\det\left(\Sigma^{-1}+\Omega^{-1}\right).$$

Taking the derivative w.r.t. $\frac{1}{\omega^2}$, we see that

$$\Gamma_{i,i}=\left(\left(\Sigma^{-1}+\text{diag}\left(\frac{1}{\omega_1},\ldots,\frac{1}{\omega_p}\right)\right)^{-1}\right)_{i,i}=\left(\Sigma-\Sigma\left(\Sigma+\Omega\right)^{-1}\Sigma\right)_{i,i} \tag{3.24}$$

has to hold for all $i=1,\ldots,(p\times(n+1))$, i.e. we have to choose ω^2 such that the posterior marginal variances $\Gamma_{i,i}$ coincide with the marginal variances of $\mathbf{G}_\psi$.

Several problems arise: First of all, Equation (3.24) is not guaranteed to have a solution. For the i-th unit-vector $e_i\in\mathbf{R}^p$ we can reformulate Equation (3.24) to

$$\Sigma_{i,i}-\Gamma_{i,i}=e_i^T\Sigma^T\left(\Sigma+\Omega\right)^{-1}\Sigma e_i>0$$

and so we require $\Gamma_{i,i}<\Sigma_{i,i}$. While the law of total covariance asserts that

$$\Sigma=\mathbb{E}\underbrace{\text{Cov}\left(X|Y\right)}_{=\Gamma}+\text{Cov}\left(\mathbb{E}\left(X|Y\right)\right),$$

it does not guarantee $\Gamma\prec\Sigma$, which would imply $\Gamma_{i,i}<\Sigma_{i,i}$.

Second, even if there is an analytical solution Ω to Equation (3.24), in the CE-method we replace $\Gamma_{i,i}$ by the observed marginal variances $\hat{\Gamma}_{i,i}$ obtained by importance sampling. The variation introduced

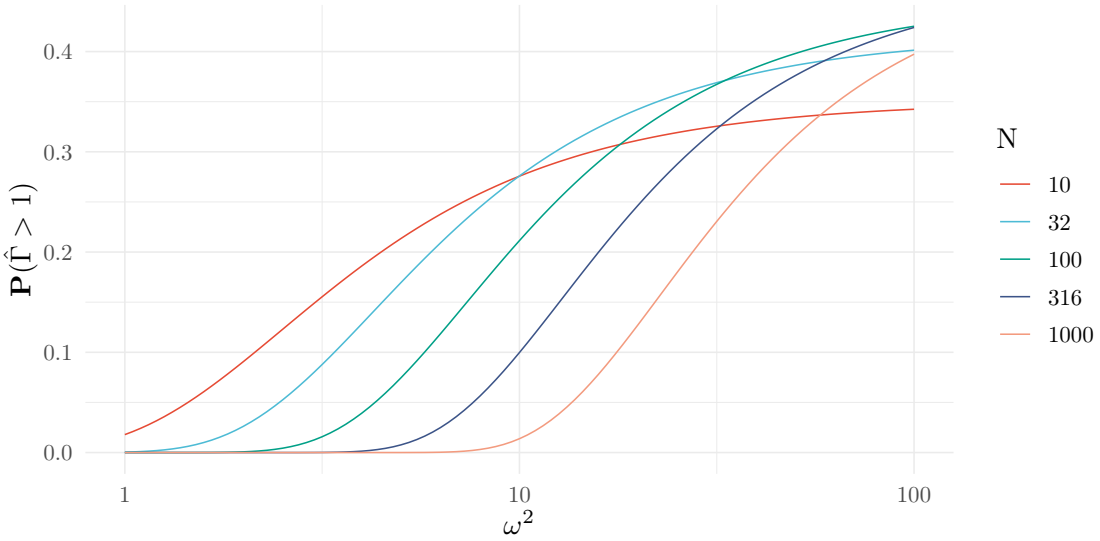

Figure 3.2: We show the probability that the estimated posterior variance $\hat{\Gamma}$ is bigger than the prior variance 1 when varying the noise variance ω^2. todo: ausführlicher beschreiben

by simulation can then lead to situations where $\hat{\Gamma}_{i,i} > \Sigma_{i,i}$. As an example take $X \sim \mathcal{N}(0,1)$, and $Y = X + \eta$ for $\eta \sim \mathcal{N}(0, \omega^2)$. Then the conditional variance of X given $Y = y$ is $\Gamma = 1 - \frac{1}{1+\omega^2}$. Given N i.i.d. samples $X^1, \dots X^N$ from this distribution, their empirical variance $\hat{\Gamma} = \frac{1}{N}\sum_{i=1}^{N}(X^i - \bar{X})^2$ follows a scaled χ^2_{N-1} distribution, i.e. $\frac{N\hat{\Gamma}}{\Gamma} \sim \chi^2_{N-1}$. Notice that we use the non-Bessel corrected version of the empirical variance here, as it is the maximum-likelihood estimate.

Then

$$\mathbf{P}\left(\hat{\Gamma} > 1\right) = \mathbf{P}\left(\frac{N\hat{\Gamma}}{\Gamma} > \frac{N}{\Gamma}\right) = 1 - F_{\chi^2_{N-1}}\left(N\left(1 + \frac{1}{\omega^2}\right)\right)$$

is the probability that Equation (3.24) has no solution $\omega^2 \in \mathbf{R}_{\geq 0}$. Here $F_{\chi^2_{N-1}}$ is the cumulative distribution function of the χ^2_{N-1} distribution. As ω^2 goes to ∞, this probability approaches $1 - F_{\chi^2_{N-1}}(N)$ which, for large N, is approximately $1 - F_{\chi^2_{N-1}}(N-1) \approx \frac{1}{2}$, as $\chi^2_{N-1} \approx \mathcal{N}\left(N-1, 2(N-1)\right)$ (Johnson, Kotz, and Balakrishnan, 1994, Section 18.5). We illustrate this in Figure 3.2, displaying the probability of failure in this setting for various combinations of N and ω^2. In this figure, we see that with growing N the threshold for ω^2 leading to non-negligible failure probability becomes larger, as expected. Thus, even in the very simple univariate Gaussian setting, for every N there is an ω^2 such that the CE-method fails for Equation (3.21) with practically relevant probability.

In higher-dimensional settings, e.g. when applying the CE-method to SSMs, we can expect this phenomenon to occur even more often. In the extreme case of independent marginals, i.e. when Σ is a diagonal matrix, Equation (3.24) reduces to $(n+1)p$ many decoupled equations, where $\hat{\Gamma}_{i,i}, i = 1, \dots, (n+1)p$ are independent. If all $q_i = \mathbf{P}\left(\Gamma_{i,i} > \Sigma_{i,i}\right)$ are identical to $q \in (0,1)$, e.g. because Σ and Ω are multiples of the identity, the number of failures follows a Binom $\left((n+1)p, q\right)$ distribution, so that even small q may lead to a non-negligible number of failures if the number of observations is high.

Finally, in the multivariate setting, the system (3.24) has no analytical solution. Instead, we have to resort to numerical methods to find a solution Ω. Unfortunately, even evaluating the right-hand side of (3.24) requires $\mathcal{O}(m^3)$ operations, as we have to invert $\Sigma + \Omega$. Additionally, we cannot hope to reuse a singular-value, LR, or eigenvalue-decomposition for further evaluations, as Σ and Ω are not guaranteed to be jointly diagonalizable. In the SSM context we may use the Kalman-smoother to compute the marginal variances, but have to re-run the smoother for every evaluation.

If we admit noise variance ∞ in the univariate setting, then $\Gamma > 1$ implies that the CE-method

chooses this as the estimate, i.e. $\mathbf{G}_{\hat{\psi}_{\text{CE}}}$ is $\mathcal{N}(0,1)$, which is equal to the prior. We can interpret this as having a missing observation, which, going back to the SSM context, the Kalman-filter (Algorithm 1) can handle with only simple modifications, see e.g. (Durbin and Koopman, 2012, Section 4.10). However, if there are a lot of failures, the optimally chosen $\mathbf{G}_{\hat{\psi}_{\text{CE}}}$ will be close to the prior distribution of states X, and importance sampling is unlikely to be effective. Hence, we turn to another approach that allows us to apply the CE-method to SSMs.

3.6.2 The Markov-approach

An alternative family of Gaussian proposals is given by directly modeling a Gaussian Markov process on the states $X_{:n}$. Again, this is sensible given the Markov structure of the target. This parametrization is more flexible than using the posterior of a GLSSM with fixed prior as the proposal. This flexibility, however, comes at the cost of requiring a larger number of parameters. Here we propose with $\mathbf{G}_\psi$ where

$$\begin{aligned}
\mathbf{G}_\psi &= \mathcal{L}(U+v),\\
v &\in \mathbf{R}^{(n+1)m},\\
U_0 &\sim \mathcal{N}(0, R_0 R_0^T),\\
U_t &= C_t U_{t-1} + R_t \nu_t,\\
C_t &\in \mathbf{R}^{m\times m},\\
\nu_t &\sim \mathcal{N}(0, I_m),\\
R_t &\in \mathbf{R}^{m\times m} \text{ lower triangular with positive diagonal}
\end{aligned} \tag{3.25}$$

for $t = 1, \dots, n$, with U_0 and $\nu_1, \dots, \nu_n$ independent. The number of parameters in

$$\psi = (v, C_1, \dots, C_n, R_0, \dots, R_n)$$

is $(n+1)\cdot m$ for the mean v, $n \cdot m^2$ for the transition matrices C_t and $(n+1)\frac{m(m-1)}{2}$ for the Cholesky roots of innovation covariances, totaling $\mathcal{O}(n \cdot m^2)$ many parameters. While these are considerably more parameters than for the GLSSM-approach for large state dimension m, we will see in the later part of this section that finding the optimal parameters for the CE-method can be done analytically.

This approach, which we term the **Markov-approach**, was originally proposed by (Richard and Zhang, 2007) for general unnormalized transition kernels as EIS proposals. However, because of its lower number of parameters, one should favor the GLSSM-approach for EIS that operates on the signals, see (Koopman, Lit, and Nguyen, 2019).

To perform importance sampling with $\mathbf{G}_\psi$ in model (3.25) we not only need to simulate from $\mathbf{G}_\psi$ but also evaluate the unnormalized importance sampling weights $w(x) = \frac{p(x|y)}{g_\psi(x)}$. Simulation from $\mathbf{G}_\psi$ is achieved by a simple recursion. For the weights note that

$$w(x) \propto \frac{p(y|x)p(x)}{g_\psi(x)} = \prod_{t=0}^{n} \frac{p(y_t|x_t)p(x_t|x_{t-1})}{g_\psi(x_t|x_{t-1})}, \tag{3.26}$$

where $p(x_0|x_{-1}) = p(x_0)$ and $g_\psi(x_0|x_{-1}) = g_\psi(x_0)$.

The Markov structure of model (3.25) implies that the precision matrix of $\mathbf{G}_\psi$ is sparse, i.e. it has a block-tridiagonal form. This is a well-known property of the precision matrix of Gaussian random vectors, as the following two classical lemmas show. We show their proofs here for completeness. For a general treatment, we refer the reader to (Lauritzen, 1996, Chapters 3 and 5).

Lemma 3.10. *Let (X, Y) be jointly Gaussian with distribution $\mathcal{N}(\mu, \Sigma)$ where*

$$\mu = (\mu_X, \mu_Y)$$

and

$$\Sigma = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix},$$

are partitioned according to the dimensions of X and Y and Σ is non-singular. If

$$P = \Sigma^{-1} = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{XY} & \Sigma_{YY} \end{pmatrix}^{-1} = \begin{pmatrix} P_{XX} & P_{XY} \\ P_{YX} & P_{YY} \end{pmatrix}$$

is the precision matrix of (X, Y), partitioned as is Σ, then $\mathrm{Cov}(X|Y) = P_{XX}^{-1}$.

Proof. Without loss of generality, assume that both X and Y are centered. The conditional density $p(x|y)$ is proportional (in x) to the joint density $p(x, y)$ with

$$\log p(x, y) = -\frac{1}{2} \begin{pmatrix} x & y \end{pmatrix} P \begin{pmatrix} x \\ y \end{pmatrix} + C = -\frac{1}{2} \left(x^T P_{XX} x + 2 x^T P_{XY} y \right) + C',$$

for constants C, C' that do not depend on x. As the conditional distribution of X given $Y = y$ is Gaussian (by Lemma 3.1), its covariance matrix is P_{XX}^{-1}. □

Lemma 3.11. *Let $(X, Y, Z) \sim \mathcal{N}(\mu, \Sigma)$ be jointly Gaussian with non-singular Σ. Then $X \perp Y|Z$ if, and only if, the sub-matrix of the precision matrix $P = \Sigma^{-1}$ whose rows correspond to the entries of X and columns correspond to the entries of Y is the 0 matrix.*

Proof. Partition the conditional covariance matrix into

$$\mathrm{Cov}\left((X, Y)|Z\right) = \begin{pmatrix} \Sigma_{XX|Z} & \Sigma_{XY|Z} \\ \Sigma_{YX|Z} & \Sigma_{YY|Z} \end{pmatrix}.$$

As all distributions involved are Gaussian, $X \perp Y|Z$ is equivalent to $\mathrm{Cov}\left((X, Y)|Z\right)$ being a block-diagonal matrix with blocks $\Sigma_{XX|Z}$ and $\Sigma_{YY|Z}$, which is equivalent to its inverse being a block-diagonal matrix with blocks $\Sigma_{XX|Z}^{-1}$ and $\Sigma_{YY|Z}^{-1}$. Its inverse is, by Lemma 3.10, the sub-matrix of P whose rows and columns correspond to X and Y. □

Applying Lemma 3.11 to model (3.25), we see that its precision matrix P is sparse, i.e. it is a block-tri-diagonal matrix, as $U_t \perp U_s|U_{-t,-s}$ for $|t - s| > 1$ and $U_{-t,-s}$ being the vector of all $U_0, \ldots, U_n$ except for U_t, U_s. Thus, the only entries of P that are potentially non-zero are those whose row and column correspond to (U_t, U_t) for $t = 0, \ldots, n$, (U_t, U_{t-1}) and (U_{t-1}, U_t) for $t = 1, \ldots, n$. Therefore, P has the following block-tridiagonal structure:

$$P = \begin{pmatrix} P_{0,0} & P_{0,1} & 0 & \cdots & \cdots & 0 & 0 \\ P_{1,0} & P_{1,1} & P_{1,2} & 0 & \cdots & 0 & 0 \\ 0 & P_{2,1} & P_{2,2} & P_{2,3} & \cdots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 & 0 \\ 0 & 0 & 0 & \cdots & P_{n-1,n-2} & P_{n-1,n-1} & P_{n-1,n} \\ 0 & 0 & 0 & \cdots & 0 & P_{n,n-1} & P_{n,n} \end{pmatrix}. \tag{3.27}$$

As the precision matrix is the natural parameter for the multivariate Gaussian exponential family, we see that model (3.25), parameterized by $(P^{-1}v, P)$ form a natural exponential family and we can apply Theorem 3.6 to obtain a central limit theorem when applying the CE-method for this model.

The sparsity of P implies that $P = LL^T$ has a sparse Cholesky root L, which will make computations efficient. To see that L is sparse, we apply the following Theorem, slightly adapted to our notation, from the theory of Gaussian-Markov-Random-fields (GMRF), i.e. Gaussian models whose dependency structure is given by a graph, with edges between nodes indicating non-zero entries in the precision matrix.

Theorem 3.10 ((Gelfand et al., 2010, Theorem 12.14))**.** *Let $X = (X_0, \ldots, X_n) \in \mathbf{R}^{(n+1)m}$ be a GMRF wrt to the labeled graph G, with mean μ and symmetric positive-definite precision matrix P. Let L be the Cholesky factor of P and define for $0 \leq t < s \leq n$ the future of t except s as*

$$F(t, s) = \{t + 1, \ldots, s - 1, s + 1, n\}.$$

Then

$$X_t \perp X_s | X_{F(t,s)} \Leftrightarrow L_{t,s} = 0.$$

In the preceding theorem $X_{F(t,s)}$ is the vector of all X_u for $u \in F(t,s)$ and $L_{t,s} \in \mathbf{R}^{m\times m}$ is the sub-matrix of L whose rows correspond to X_t and columns to X_s. From Theorem 3.10 we immediately obtain the following:

Corollary 3.2 (sparsity of L in model (3.25))**.** *Let $U \sim \mathbf{G}_\psi$ as in Equation (3.25), $P \succ 0$ be the precision matrix of $\overleftarrow{U} = (U_n, \dots, U_0)$ and L be the Cholesky root of P. Then L is a lower-block-diagonal matrix, with at most $n\,m^2 + (n+1)\,m\frac{m-1}{2}$ non-zero entries:*

$$L = \begin{pmatrix} L_{n,n} & 0 & \cdots & \cdots & \cdots & 0 & 0 \\ L_{n-1,n} & L_{n-1,n-1} & 0 & \cdots & \cdots & 0 & 0 \\ 0 & L_{n-2,n-1} & L_{n-2,n-2} & 0 & \cdots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 & 0 \\ 0 & 0 & 0 & \cdots & L_{1,2} & L_{1,1} & 0 \\ 0 & 0 & 0 & \cdots & 0 & L_{0,1} & L_{0,0} \end{pmatrix}, \tag{3.28}$$

where $L_{t,t} \in \mathbf{R}^{m\times m}, t = 0, \dots, n$ are lower triangular matrices with positive diagonal entries and $L_{t-1,t} \in \mathbf{R}^{m\times m}, t = 1, \dots, n$ are square matrices.

From L in Corollary 3.2 we obtain an iterative method of sampling from $\mathbf{G}_\psi$: If $v + U \sim \mathbf{G}_\psi$, then, as $\operatorname{Cov} U = \left(LL^T\right)^{-1} = L^{-T}L^{-1}$, it holds that $L^T U \sim \mathcal{N}(0, I)$ follows a standard normal distribution. Thus to simulate from $\mathbf{G}_\psi$ we may solve

$$L^T U = \overleftarrow{Z}$$

where $\overleftarrow{Z} = (Z_n, \dots, Z_0) \sim \mathcal{N}(0, I)$. Using the structure available in L, we see that this is equivalent to first solving

$$L_{0,0}^T U_0 = Z_0$$

and then recursively solving for $t = 1, \dots, n$

$$L_{t,t}^T U_t + L_{t-1,}^T U_{t-1} = Z_{t-1}.$$

Rearranging terms, provided $L_{t,t}$ is non-singular, we end up with the Markov-process

$$U_t = L_{t,t}^{-T} L_{t-1,t}^T U_{t-1} + L_{t,t}^{-T} Z_t, \tag{3.29}$$

where Z_t is, by construction, independent of U_{t-1}. Thus for model (3.25), we obtain

$$\begin{aligned} R_t &= L_{t,t}^{-T} \text{ for } t = 0, \dots, n, \\ C_t &= L_{t,t}^{-T} L_{t-1,t}^T \text{ for } t = 1, \dots, n. \end{aligned} \tag{3.30}$$

Here we see why we chose to use $\overleftarrow{U}$ in Corollary 3.2: had we applied Theorem 3.10 to U directly we would have ended up with a Markov process in reverse time.

We now turn our attention to applying the CE-method to model (3.25). Following a similar argument as in the discussion surrounding Equation (3.23), we see that we may match the mean v to that of $\mathbf{P}$ and it suffices to choose P, the precision matrix of U, such that it minimizes

$$\frac{1}{2} \operatorname{trace}\left(P\hat{\Gamma}\right) - \frac{1}{2} \log\det P \tag{3.31}$$

where $\hat{\Gamma}$ is the importance sampling estimate of the joint covariance matrix of all states X. This is equivalent to minimizing

$$\mathcal{D}_{\mathrm{KL}}\left(\mathcal{N}(0, \hat{\Gamma}) \middle\| \mathcal{N}(0, P^{-1})\right).$$

Here P is restricted to precision matrices that may arise in model (3.25), i.e., by Corollary 3.2, $P = LL^T$ where L possess structure as in (3.28). At first glance, this problem seems more involved than solving Equation (3.24): after all, the optimal P depends on the whole covariance matrix $\hat{\Gamma}$. However, it turns out that the sparsity we enforce in L allows us to compute analytically the optimal $\hat{L}$ that minimizes Equation (3.31). Additionally, due to the Markov-structure of our proposal, $\hat{L}$ depends only on the block-tri-diagonal component of $\hat{\Gamma}$, i.e. only the covariances $\mathrm{Cov}(X_t, X_{t-1})$ and $\mathrm{Cov}(X_0)$ are required. This is sensible - all information about the Markov transitions is encoded in these covariances if we assume that X is a Gaussian Markov process.

To make this argument rigorous, let us apply the following result (stated in our notation).

Theorem 3.11 ((Schäfer, Katzfuss, and Owhadi, 2021, Theorem 2.1)). *Let Γ be a positive-definite matrix of size $n \times n$. Given a lower-triangular sparsity set $S \subset \{1, \dots, n\}^2$, i.e. $i \geq j$ for all $(i,j) \in S$, let*

$$\hat{L} = \mathrm{argmin}_{L \in \mathcal{S}} \mathcal{D}_{KL}\left(\mathcal{N}(0,\Gamma) \middle|\middle| \mathcal{N}\left(0, (LL^T)^{-1}\right)\right)$$

be the Cholesky root of the closest Gaussian (wrt. the KL-divergence) with sparsity $\mathcal{S} = \{A \in \mathbf{R}^{n\times n} : A_{i,j} \neq 0 \Rightarrow (i,j) \in S\}$.

Then the following holds: The nonzero entries of the i-th column of $\hat{L}$ are given by

$$L_{s_i,i} = \frac{\Gamma^{-1}_{s_i,s_i} e_1}{\sqrt{e_1^T \Gamma^{-1}_{s_i,s_i} e_1}}, \tag{3.32}$$

where $s_i = \{j : (i,j) \in \mathcal{S}\}$, Γ_{s_i,s_i} is the restriction of Γ to the set of indices s_i and $e_1 \in \mathbf{R}^{|s_i|}$ is the first unit vector.

Exploiting the Markov structure of our proposals, we immediately obtain the following:

Corollary 3.3. *Let $\mathcal{S}$ be the sparsity set of a Gaussian Markov process of the form Equation (3.25), i.e.*

$$\mathcal{S} = \left\{((t,i),(s,j)) \in (\{0,\dots,n\} \times \{1,\dots,m\})^2 \mid (t = s \text{ and } i \geq j) \text{ or } t = s+1\right\},$$

see also Equation (3.28), and let Γ be a positive definite matrix of size $((n+1)m) \times (n+1)m$ with blocks

$$\Gamma_{s,t} = (\Gamma_{(s,i),(t,j)})_{i,j=1,\dots,m}.$$

Then $\hat{L}$ in Theorem 3.11 depends only on the block-diagonal entries $\Gamma_{t,t}$, $t = 0,\dots,n$ and block off-diagonal entries $\Gamma_{t,t+1}$, $t = 0,\dots,n$.

If, in particular, Γ is the covariance matrix of Gaussian Markov process, $\hat{L} = \mathrm{chol}(\Gamma^{-1})$.

We have thus shown the following: The covariance matrix of the KL-optimal Gaussian Markov process for the positive definite covariance matrix Γ with $\mathcal{O}(n^2m^2)$ entries only depends on $\mathcal{O}(nm^2)$ many entries, the marginal covariances. In particular, if we can find a centered Gaussian Markov process $(X_t)_{t=0,\dots,n}$ whose marginal covariances fulfill

$$\begin{aligned} \mathrm{Cov}(X_t) &= \Gamma_t && t = 0,\dots,n \\ \mathrm{Cov}(X_t, X_{t+1}) &= \Gamma_{t,t+1} && t = 0,\dots,n, \end{aligned}$$

then its law $\mathcal{L}(X)$ is the one we seek. The following proposition puts all the pieces together.

Proposition 3.6 (the CE-method for the Markov proposal). *Let $\mathbf{P}$ be a probability measure on $\mathbf{R}^{(n+1)\times m}$ with mean μ and positive definite covariance matrix Γ, partitioned into blocks*

$$\Gamma_{s,t} = (\Gamma_{(s,i),(t,j)})_{i,j=1,\dots,m}.$$

Let

$$\begin{pmatrix} J_{t,t} & 0 \\ J_{t+1,t} & Z_{t+1,t+1} \end{pmatrix} = \mathrm{chol} \begin{pmatrix} \Gamma_{t,t} & \Gamma_{t,t+1} \\ \Gamma_{t+1,t} & \Gamma_{t+1,t+1} \end{pmatrix}.$$

Then the optimal cross-entropy parameter

$$\psi_{CE} = \operatorname{argmin}_{\psi=(v,C_1,\dots,C_n,R_0,\dots,R_n)} \mathcal{D}_{KL}\left(\mathbf{P}||\mathbf{G}_\psi\right)$$

adapt notation to model? use tildes as cholesky roots

for the Markov proposal $\mathbf{G}_\psi$ *from model* (3.25) *exists and is unique. The components of* ψ_{CE} *are given by*

$$\begin{aligned} v &= \mu \\ R_0 &= \operatorname{chol}(\Gamma_{0,0}) \end{aligned}$$

and for $t = 1, \dots, n$

$$\begin{aligned} C_t &= J_{t+1,t} J_{t,t}^{-1} \\ R_t &= Z_{t+1,t+1} \end{aligned}$$

Thus, given ν *and* Γ, ψ_{CE} *can be obtained in* $\mathcal{O}(n\, m^3)$ *many operations.*

Proof. It only remains to show the uniqueness and existence of ψ_{CE}, as well as its representation. The discussion surrounding Equation (3.31) shows that $v = \mu$ has to hold, so we may assume that $\mathbf{P}$ and the proposal are both centered. As Γ is positive definite, so are all of its sub-matrices, and we may apply Corollary 3.3. Therefore, if we can show that there is a unique Gaussian Markovian probability measure whose covariance matrix matches Γ as in that corollary we are done.

Let $(U_t)_{t=0,\dots,n} \sim \mathbf{G}_{\psi_{\mathrm{CE}}}$. Then

$$\operatorname{Cov}(U_0) = R_0 R_0^T = \Gamma_{0,0},$$

and from the Cholesky decomposition we obtain for $t = 0, \dots, n-1$

$$\begin{pmatrix} J_{t,t} J_{t,t}^T & J_{t,t} J_{t+1,t}^T \\ J_{t+1,t} J_{t,t}^T & J_{t+1,t} J_{t+1,t}^T + Z_{t+1,t+1} Z_{t+1,t+1}^T. \end{pmatrix} = \begin{pmatrix} \Gamma_{t,t} & \Gamma_{t,t+1} \\ \Gamma_{t+1,t} & \Gamma_{t+1,t+1} \end{pmatrix}.$$

As $Z_{t+1,t+1}$ is a lower triangular matrix with positive diagonal and

$$\Gamma_{t+1,t+1} - \Gamma_{t+1,t} \Gamma_t^{-1} \Gamma_{t,t+1} = Z_{t+1,t+1} Z_{t+1,t+1}^T,$$

it is the Cholesky root of the Schur complement $\Gamma_{t+1,t+1} - \Gamma_{t+1,t}\Gamma_t^{-1}\Gamma_{t,t+1}$, which, recalling Lemma 3.1, we can think of as a conditional covariance matrix. Therefore, using induction over $t = 0, \dots, n-1$, we obtain

$$\begin{aligned} \operatorname{Cov}\left(U_{t+1}\right) &= C_{t+1} \operatorname{Cov}\left(U_t\right) C_{t+1}^T + R_{t+1} R_{t+1}^T \\ &= J_{t+1,t} J_{t,t}^{-1} \Gamma_{t,t} J_{t,t}^{-T} J_{t+1,t}^T + \Gamma_{t+1,t+1} - \Gamma_{t+1,t} \Gamma_t^{-1} \Gamma_{t,t+1} \\ &= \Gamma_{t+1,t+1} \end{aligned}$$

and

$$\operatorname{Cov}\left(U_{t+1}, U_t\right) = C_t \operatorname{Cov}\left(U_t\right) = J_{t+1,t} J_{t,t}^{-1} J_{t,t} J_{t,t}^T = \Gamma_{t+1,t}.$$

This shows the existence. For uniqueness, note that model (3.25) enforces that R_t is a lower triangular matrix with positive diagonals. As $R_{t+1} R_{t+1}^T$ is the conditional covariance of U_{t+1} given U_t which is, by Lemma 3.1 given by $\Gamma_{t+1,t+1} - \Gamma_{t+1,t}\Gamma_{t,t}^{-1}\Gamma_{t+1,t}$. Thus the R matrices are unique as well. As $\operatorname{Cov}(U_{t+1}, U_t) = C_t \operatorname{Cov}(U_t)$, we can show that, additionally, also C_t is unique. □

rewrite everything in terms of this prop

When using the CE-method, we do not have access to the mean and covariances necessary to apply this proposition. Instead, we may apply the CE-method to estimate ψ in model (3.25) by replacing

these unknown moments with their importance sampling estimates. Given importance samples $U^1, \dots, U^N$ for $\mathcal{L}(X|Y=y)$ and associated auto-normalized weights $W^1, \dots, W^N$, we estimate v by

$$\hat{v} = \sum_{i=1}^{N} W^i X^i \tag{3.33}$$

and the empirical covariance matrices

$$\begin{aligned} \widehat{\mathrm{Cov}}\,(X_t, X_{t-1}) &= \sum_{i=1}^{N} W^i (X^i_{t:t-1} - \hat{v}_{t-1:t})(X^i_{t:t-1} - \hat{v}_{t-1:t})^T \\ \widehat{\mathrm{Cov}}\,(X_0) &= \sum_{i=1}^{N} W^i (X^i_0 - \hat{v}_0)(X^i_0 - \hat{v}_0)^T \end{aligned} \tag{3.34}$$

These steps are summarized in Algorithm 7.

Algorithm 7 The CE-method for the Markov proposal (3.25)

Require: LCSSM (Definition 3.5), observations Y, initial estimate $\hat{\psi}^0 = \left(v^0, C^0, R^0\right)$, sample size N

1: set $l = 0$
2: **repeat**
3: sample $U^1 + v^l, \dots, U^N + v^l \overset{\text{i.i.d}}{\sim} \mathbf{G}_{\hat{\psi}^l}$ with fixed seed ▷ Equation (3.25)
4: determine auto-normalized weights $W^1, \dots, W^N$ ▷ Equation (3.26)
5: estimate $\hat{v}^{l+1}$ ▷ Equation (3.33)
6: estimate $\widehat{\mathrm{Cov}}(U_t, U_{t-1}), t = 1, \dots, n$, and $\widehat{\mathrm{Cov}}(U_0)$ ▷ Equation (3.34)
7: determine C^{l+1} and R^{l+1} ▷ Proposition 3.6
8: set $\hat{\psi}^{l+1} = \left(\hat{v}^{l+1}, C^{l+1}, R^{l+1}\right)$
9: set $l = l + 1$
10: **until** $\hat{\psi}^l$ converged
11: **return** $\hat{\psi}_{\mathrm{CE}} = \hat{\psi}^l$

To run Algorithm 7 we require an initial value for $\hat{\psi}^0$. If a suitable $\hat{\psi}^0$ is not available, we can obtain one from the LA by sampling $X^1, \dots, X^N$ from the LA and performing steps 5 to 8 from the loop. Alternatively, we could also directly base our initial value on the smoothing distribution of the GLSSM that the LA is based on. The Kalman smoother (Algorithm 2) provides us with the analytically available covariances $\mathrm{Cov}\,(X_t, X_{t-1}|Z=z)$ and the marginal covariance $\mathrm{Cov}\,(X_0|Z=z)$ can be computed as well.

The convergence criteria in Algorithm 7 is similar to that used for EIS: we stop until the absolute or entry-wise relative difference of $\hat{\psi}^l$ and $\hat{\psi}^{l+1}$ is smaller than a predetermined threshold, or a fixed number of iterations has passed. For the matrices involved, we use the Frobenius norm and the Euclidean distance for the mean v.

In Line 3 we use the standard praxis of CRNs to ensure numerical convergence. This is similar to EIS and the maximum likelihood estimates from Section 3.7.

We give an overview of the time and space complexities of each line in Algorithm 7 in Table 3.1 The total time complexity of a single iteration of Algorithm 7 is $\mathcal{O}\left(N\,n\,m^2 + n\,m^3\right)$ and its space complexity is $\mathcal{O}\left(N\,n\,m + n\,m^2\right)$. Let us elaborate on the complexities of each step:

Line 3 Generate N i.i.d. samples from model (3.25), where each simulation requires $\mathcal{O}(n)$ matrix-vector multiplications of dimension m.

Line 4 To evaluate the weights, Equation (3.26), we have to evaluate for every sample $\mathcal{O}(n)$-times the density of a m-variate Gaussian distribution, while this usually has time-complexity $\mathcal{O}(m^3)$, we have access to the Cholesky root R_t, so this step has only time-complexity $\mathcal{O}(m^2)$.

step	time complexity	space complexity
simulation (Line 3)	$\mathcal{O}\left(N\,n\,m^2\right)$	$\mathcal{O}\left(N\,n\,m\right)$
weights (Line 4)	$\mathcal{O}(N\,n\,m^2)$	$\mathcal{O}\left(N\right)$
estimating v (Line 5)	$\mathcal{O}(N\,n\,m)$	$\mathcal{O}\left(n\,m\right)$
estimating covariances (Line 6)	$\mathcal{O}(N\,n\,m^2)$	$\mathcal{O}\left(n\,m\right)$
determining C and R (Line 7)	$\mathcal{O}(n\,m^3)$	$\mathcal{O}(n\,m^2)$

Table 3.1: Time and space complexities of individual steps in Algorithm 7.

In Equation (3.26) we also need to compute $p(y_t|x_t)$ and $p(x_t|x_{t-1})$. Assuming conditional independence of observations, $p(y_t|x_t) = \prod_{i=1}^m p(y_t^i|(B_t x_t)^i)$, evaluating the first term requires only $\mathcal{O}(m^2)$ operations. For the second term, if we allow pre-computation of the Cholesky roots of innovations off-line (in $\mathcal{O}(m^3)$ time), this step reduces to $\mathcal{O}(m^2)$ as well.

Line 5 Calculating the weighted mean $\hat{v} \in \mathbf{R}^{(n+1)m}$, Equation (3.33), requires $\mathcal{O}(N\,n\,m)$ operations.

Line 6 Calculating the weighted covariance matrices, Equation (3.34), requires $(n+1)$ times multiplying N many $m \times 1$ with $1 \times m$ vectors.

Line 7 For each of the $\mathcal{O}(n)$ many C_t and R_t we have to calculate Cholesky decompositions and invert triangular matrices of dimension m.

An efficient implementation of Algorithm 7 can improve on the practically relevant computational time. There is no need to calculate the C_t matrices explicitly, instead we can calculate $C_t U_{t-1} = J_{t+1,t} J_{t,t}^{-1} U_{t-1}$ efficiently by back-substitution, as $J_{t,t}$ is a lower triangular matrix.

The main bottleneck for space lies in the $\mathcal{O}(N\,n\,m)$ simulation part, and we may reduce this by simulating twice from model (3.25) using CRNs, and only storing the samples for a single time step (dimension $\mathcal{O}(N\,m)$) in each simulation. In the first pass, we only calculate the weights, and in the second pass, we calculate $\hat{v}$ and the required covariance matrices. For this, we only need the $2N$ samples of dimension m from time t and $t+1$, i.e. $\mathcal{O}(N\,m)$ space. This reduces the total space complexity to $\mathcal{O}(N\,m + n\,m^2)$.

We demonstrate these improvements in Algorithm 8. Additionally, we calculate the weights on the log scale for numerical stability.

The advantage of Algorithms 7 and 8 over applying the CE-method to the GLSSM model (3.21) are multiple: First of all, as long as the involved covariance matrices are positive definite, the two algorithms produce valid proposals, i.e. they do not have the degeneracy problem we observed in Section 3.6.1. When matrices are only positive-semi definite, replacing inverses with generalized inverses still yields a valid model. Additionally, determining the optimal parameters (v, C, R) or (v, J, R) is numerically stable, involving only inversion of small matrices. Compare this with solving Equation (3.24), where we need to employ a numerical scheme to solve for the diagonal entries of Ω.

After having determined $\hat{\psi}_{\mathrm{CE}}$ for model (3.25), generating N samples requires only $\mathcal{O}(N\,n\,m^2)$ operations, whereas sampling from model (3.21) requires $\mathcal{O}(n\,m^3 + N\,n\,m^2)$ operations, as we need an initial run of the Kalman filter. Unless $N < m$, this difference is negligible, and the case where $N < m$ is not really of interest, as we would expect importance sampling to require a much larger number of samples, i.e. $N \gg m$.

However, the two algorithms presented in this section also come with some drawbacks, especially if the dimension m of states is large. This affects the algorithms in multiple ways: when m is large, computation of the Cholesky decomposition in Proposition 3.6 becomes more time-intensive. Additionally, the dimension of the parameter ψ increases quadratically in m, so we expect convergence to be slower, requiring a larger sample size N to find the optimal $\hat{\psi}_{\mathrm{CE}}$. For an empirical study in this direction, see Section 3.8.

Algorithm 8 Time and space improved version of Algorithm 7. Instructions involving the free index i are to be performed for all $i = 1, \dots, N$ samples. For simplicity of notation we let $R^l = (R_0^l, \dots, R_n^l)$ and $J^l = (J_{0,0}^l, J_{1,0}^l, \dots, J_{n-1,n-1}^l, J_{n,n-1}^l)$ for $l \in \mathbf{N}_0$.

Require: LCSSM (Definition 3.5), observations Y, initial estimate $\hat{\psi}^0 = \left(v^0, R^0, J^0\right)$, sample size N

set $l = 0$
repeat
  simulate $\nu_0^1, \dots, \nu_0^N \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, I)$
  set $U_0^i = R_0^l \nu_0^i$
  set $X_0^i = v_0^l + U_0^i$
  set $\log w^i = \log p(y_0|X_0^i) + \log p(X_0^i) + \frac{1}{2}\|\nu_0^i\|^2$ ▷ $\log g(X_0^i) = -\frac{1}{2}\|\nu_0^i\|_2^2 + C$
  store current RNG state
  for $t \leftarrow 1, \dots, n$ **do**
    simulate $\nu_t^1, \dots, \nu_t^N \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, I)$
    set $U_t^i = (J_{t+1,t})^T (J_{t,t})^{-1} U_{t-1}^i + R_t^l \nu_t^i$ ▷ backsubstitution
    set $X_t^i = v_t^l + U_t^i$
    set $\log w^i = \log w^i + \log p(y_t|X_t^i) + \log p(X_t^i|X_{t-1}^i) + \frac{1}{2}\|\nu_t^i\|^2$
  end for
  set $\log w^i = \log w^i - \max_{i=1,\dots,N} \log w^i$ ▷ ensure $\log w^i \leq 0$
  set $w^i = \exp(\log w^i)$
  set $W^i = \frac{w^i}{\sum_{i=1}^N w^i}$ ▷ auto-normalized weights
  set $v_0^{l+1} = \sum_{i=1}^N W^i X_0^i$
  restore RNG state
  for $t \leftarrow 1, \dots, n$ **do**
    simulate $\nu_t^1, \dots, \nu_t^N \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, I)$
    set $U_t^i = (J_{t+1,t})^T (J_{t,t})^{-1} U_{t-1}^i + R_t^l \nu_t^i$ ▷ backsubstitution
    set $X_t^i = v_t^l + U_t^i$
    calculate $\hat{v}_t^{l+1}$ ▷ Equation (3.33)
    calculate covariances ▷ Equation (3.33)
  end for
  set $\hat{\psi}^{l+1} = \left(\hat{v}^{l+1}, \hat{R}^{l+1}, \hat{J}^{l+1}\right)$
  set $l = l + 1$
until $\hat{\psi}^l$ converged
return $\hat{\psi}_{\text{CE}} = \hat{\psi}^l$

3.7 Maximum likelihood estimation in SSMs

Until now, we have assumed that the SSM under consideration is completely known, i.e. we have access to the true transition and observation kernels. For the models considered in this thesis (Chapter 4), this is unrealistic, as they are not based on concrete physical processes but are rather statistical approximations of the true underlying dynamics. The transition densities of, e.g., Equation (3.4) will depend on the covariance matrix of innovations, of which we have no a priori knowledge and for negative binomially distributed observations the overdispersion parameter r will be unknown. Let us denote by $\theta \in \mathbf{R}^l$ the vector of these hyperparameters.

check l / k with psis

To make this dependence explicit, we will introduce subscripts θ where appropriate, i.e. $\mathbf{P}_\theta$ is a target distribution that additionally depends on θ, p_θ its density et cetera. This section is loosely based on (Durbin and Koopman, 2012, Chapter 7 & 11) and (Chopin and Papaspiliopoulos, 2020, Chapter 14)

To determine a suitable value of θ, multiple options are available. Here, we opt for a frequentist approach, using maximum likelihood estimation to determine an optimal $\hat{\theta}$. Therefore, given observations $y \in \mathbf{R}^{(n+1)\times p}$, $\hat{\theta}$ maximizes the likelihood $p_\theta(y)$ and can be obtained as the global maximum of the following optimization problem:

$$\max_{\theta\in\Theta} p_\theta(y).$$

For numerical stability, we should maximize the log-likelihood instead, i.e. solve

$$\max_{\theta\in\Theta} \log p_\theta(y). \tag{3.35}$$

Here $\Theta \subseteq \mathbf{R}^l$ is the parameter space. To solve this optimization problem using gradient ascent algorithms, we need access to both the likelihood and its derivatives. Thus, in the following, we will assume that $\theta \mapsto \log p_\theta(y)$ is sufficiently smooth, to apply these methods, i.e. it has continuous derivatives of second order.

While the Kalman-filter (Algorithm 1) allows analytical computation of this likelihood GLSSMs, in general SSMs it is numerically intractable. The reason for this is that

$$p_\theta(y) = \int p_\theta(x, y)\mathrm{d}\mu(x)$$

is a high-dimensional integral, which is hard to evaluate numerically. Instead, we will use importance sampling to estimate the likelihood. For this, let us regard $p_\theta(x, y)$ as an unnormalized density in x. The missing integration constant is then just $p_\theta(y)$ and the normalized density is $p_\theta(x|y)$. If $\mathbf{G} \gg \mathbf{P}$ is a proposal distribution whose density g with respect to μ we can evaluate analytically, i.e. not only up to a constant, we see that for the unnormalized weights $\tilde{w}_\theta(x) = \frac{p_\theta(x,y)}{g(x)}$, that $p_\theta(y) = \mathbf{G}[\tilde{w}_\theta]$. Thus we may estimate the likelihood by

$$\widehat{p_\theta(y)} = \frac{1}{N}\sum_{i=1}^{N} \tilde{w}_\theta(X^i)$$

for $X^1, \dots, X^N \overset{\text{i.i.d}}{\sim} \mathbf{G}$ and $N \in \mathbf{N}$. To evaluate the gradient, notice that as $\nabla_\theta p_\theta(x, y) = p_\theta(x, y)\nabla_\theta \log p_\theta(x, y)$, we have, provided we can exchange integration and differentiation,

$$\begin{aligned}\nabla_\theta p_\theta(y) &= \nabla_\theta \int p_\theta(x, y)\,\mathrm{d}\mu(x) = \int p_\theta(x, y)\nabla_\theta \log p_\theta(x, y)\,\mathrm{d}\mu(x)\\ &= \mathbf{G}[\tilde{w}_\theta \nabla_\theta \log p_\theta(x, y)],\end{aligned}$$

and so we may estimate the gradient by

$$\widehat{\nabla_\theta p_\theta(y)} = \frac{1}{N}\sum_{i=1}^{N} \tilde{w}_\theta(X^i)\nabla_\theta \log p_\theta(X^i, y)$$

Similarly, we can estimate the log-likelihood by Plug-In

$$\widehat{\log p_\theta(y)} = \log\left(\frac{1}{N}\sum_{i=1}^{N}\tilde{w}_\theta(X^i)\right) \tag{3.36}$$

and its gradient, using the fact that the gradient of $\log f$ for $f : \mathbf{R}^l \to \mathbf{R}$ is $\frac{1}{f}\nabla_\theta f$, by

$$\begin{aligned}\widehat{\nabla_\theta \log p_\theta(y)} &= \left(\frac{1}{N}\sum_{i=1}^{N}\tilde{w}_\theta(X^i)\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\tilde{w}_\theta(X^i)\nabla_\theta \log p_\theta(X^i, y)\right)\\ &= \sum_{i=1}^{N} W_\theta^i \nabla_\theta \log p_\theta(X^i, y)\end{aligned}$$

where $W_\theta^i = \frac{\tilde{w}_\theta(X^i)}{\sum_{i=1}^N \tilde{w}_\theta(X^i)}$ are the auto-normalized weights. Note that, by Jensen's inequality, these estimates are biased.

To solve the optimization problem (3.35) we will again employ CRNs. If the densities involved are twice differentiable, this device ensures that the random objective function $\theta \mapsto \sum_{i=1}^N \tilde{w}_\theta(X^i)$ is twice differentiable, and so we can indeed apply gradient ascent to find a local maximum. This is an advantage of performing global importance sampling over SMC, i.e. particle filter, methods. To avoid collapse to a single particle, SMC methods perform intermediate resampling steps, which make the objective function discontinuous. While particle smoothing methods can mitigate this problem, they are more expensive than standard SMC and, as the importance sampling estimates of the log-likelihood and its gradient are biased, the usual requirements for stochastic approximation methods are not fulfilled. For a more thorough discussion of the challenges maximum likelihood estimation with SMC methods faces, we recommend (Chopin and Papaspiliopoulos, 2020, Chapter 14).

While MLEs have a strong frequentist foundation, let us stress that, for the models that we investigate in Chapter 4, the frequentist properties of the estimates are not of interest. The reason for this is that a frequentist interpretation requires us to imagine, at least hypothetically, an infinite repetition of the data-generating process. For the data at hand, such repetition is nonsensical: the pandemic is a „one-off“ event that will not be replicated under even approximately similar circumstances. Therefore, we will choose to view the estimation procedure more as a hyper-parameter tuning step, rather than true frequentist inference. While we can compute asymptotic confidence intervals for $\hat{\theta}$, see, e.g., (Durbin and Koopman, 2012, Chapter 11.6), (Chopin and Papaspiliopoulos, 2020, Chapter 14.8), these are not of practical interest for similar reasons.

As an alternative to modeling θ as fixed, but unknown, and performing maximum-likelihood estimation to obtain $\hat{\theta}$, one might also model θ as random with prior density $p(\theta)$, such that the full model becomes $p(x, y, \theta) = p(x, y|\theta)p(\theta)$. In this setup, sometimes called the Bayesian treatment of SSMs (Durbin and Koopman, 2012, Section 13.1), the main interest still lies in the posterior density $p(x, \theta|y)$, which, depending on the model at hand, can drastically increase the difficulty of the problem: even if $p(x, y|\theta)$ is an analytically tractable model such as a GLSSM, unless the prior is chosen to be conjugate, one has to resort to, e.g., MCMC-methods.

By the structure of the model, Equation (3.2), the log density and its gradient can be computed efficiently by

$$\begin{aligned}\log p_\theta(x, y) &= \log p_\theta(x_0) + \sum_{t=1}^{n} \log p_\theta(x_t|x_{t-1}) + \log p_\theta(y_t|x_t, y_{t-1})\\ \nabla_\theta \log p_\theta(x, y) &= \nabla_\theta \log p_\theta(x_0) + \sum_{t=1}^{n} \nabla_\theta \log p_\theta(x_t|x_{t-1}) + \nabla_\theta \log p_\theta(y_t|x_t, y_{t-1}),\end{aligned}$$

respectively.

Similarly, when proposing with a GLSSM or Markov-proposal for a PGSSM, the weights have similar structure, seeEquations (3.22) and (3.26), which makes calculation of $\tilde{w}$ efficient.

For the remainder of this section, let us consider the GLSSM-proposal obtained by EIS for a PGSSM with linear signal, as this is the main setting of Chapter 4. For this we obtain

$$\tilde{w}_\theta(x) = \tilde{w}_\theta(s)g(z)\frac{p_\theta(y|s)}{g(z|s)} = g(z)\prod_{t=0}^{n}\frac{p_\theta(y_t|s_t)}{g(z_t|s_t)},$$

where $s_t = B_t x_t$, $t = 0, \dots, n$, is the signal, and so the log-likelihood is given by

$$\log p_\theta(y) = \log g_\theta(z) + \log \mathbb{E}\left(w_\theta(S)|Y = y\right) \tag{3.37}$$

and can be estimated by

$$\widehat{\log p_\theta(y)} = \log g_\theta(z) + \log\left(\frac{1}{N}\sum_{i=1}^{N}\prod_{t=0}^{n}\frac{p_\theta(y_t|S_t^i)}{g(z_t|S_t^i)}\right). \tag{3.38}$$

Notice that $\log g_\theta(z)$ is the likelihood in a GLSSM, which can be computed efficiently by the standard Kalman filter (Algorithm 1). As in the GLSSM-approach we propose with an GLSSM whose state density $g(x)$ and observation matrices B_t, $t = 0, \dots, n$ are equal to those of the target, the log-likelihood $\log g_\theta(z)$ also depends on θ. The estimated gradient of the log-likelihood is

$$\widehat{\nabla_\theta \log p_\theta(y)} = \nabla_\theta \log g_\theta(z) + \sum_{i=1}^{N} W_\theta^i \sum_{t=0}^{n} \nabla_\theta \log p_\theta(y_t|S_t^i).$$

The gradient of the GLSSM log-likelihood can be obtained either numerically or analytically by employing the Kalman filter and smoother (Koopman and Shephard, 1992), however, numerical evaluation may be faster if the dimension of θ is small compared to the length of the time series, as evaluating the likelihood only requires a single application of the Kalman filter.

As the observation densities $g(z_t|s_t)$ do not depend on θ, their derivatives do not appear in the above estimate. However, when using EIS to determine an optimal proposal, the parameter $\psi = (z, \omega)$ implicitly depends on θ. Accounting for this yields the gradient

$$\widehat{\nabla_\theta \log p_\theta(y)} = \nabla_\theta \log g_\theta(z) + \sum_{i=1}^{N} W_\theta^i \left(\sum_{t=0}^{n} \nabla_\theta \log p_\theta(y_t|S_t^i) - \nabla_\theta \log g_\theta(z_t|S_t^i)\right),$$

as $\nabla_\theta \frac{1}{g_\theta(z|s)} = -\frac{1}{g_\theta(z|s)}\nabla_\theta \log g_\theta(z|s)$. The computation of this additional term is much more involved, as the parameters z, Ω are found through an iterative numerical scheme. Instead, we favor numerical differentiation of the whole procedure to evaluate the likelihood at θ, including the method of finding an optimal importance sampling scheme.

As a single evaluation of the log-likelihood can become very expensive we want our procedure to be as efficient as possible. To this end, (Durbin and Koopman, 1997) provides several improvements to the basic algorithm if the model is a PGSSM with a linear signal. Their contributions consist of a bias correction for the log-likelihood, the use of antithetic and control variables to reduce Monte-Carlo error for importance sampling and a deterministic initialization procedure. Let us briefly summarize these ideas, adapted to our notation. As the computational gains for control variates in the presence of antithetic variables seem to be limited, we do not give the same level of detail here, for an in-depth analysis, we refer the reader to the source.

For bias reduction, a second-order Taylor series expansion shows that for $\tilde{w}_\cdot = \frac{1}{N}\sum_{i=1}^{N}\tilde{w}(X^i)$,

$$\begin{aligned}\mathbb{E}\left(\log \tilde{w}_\cdot\right) - \log \mathbf{G}\tilde{w} &= \mathbb{E}\log\left(1 + \frac{\tilde{w}_\cdot - \mathbf{G}\tilde{w}}{\mathbf{G}\tilde{w}}\right) \\ &= \frac{\tilde{w}_\cdot - \mathbf{G}\tilde{w}}{\mathbf{G}\tilde{w}} - \frac{1}{2}\left(\frac{\tilde{w}_\cdot - \mathbf{G}\tilde{w}}{\mathbf{G}\tilde{w}}\right)^2 + \mathcal{O}_p(N^{-\frac{3}{2}}),\end{aligned}$$

provided $\tilde{w} \in L^3(\mathbf{G})$. Thus, estimating the second order term by $-\frac{\hat{\sigma}^2}{2N\tilde{w}_\cdot}$, where $\hat{\sigma}^2$ is the empirical variance of the unnormalized weights, we can perform a bias reduction by estimating

$$\widehat{\log p_\theta(y)} = \log(\tilde{w}_\cdot) + \log g_\theta(z) + \frac{\hat{\sigma}^2}{2N\tilde{w}_\cdot} \tag{3.39}$$

The second improvement of (Durbin and Koopman, 1997) is the use of antithetic variables and control variates, a device to reduce Monte-Carlo variance. The main idea of an antithetic variable is to construct for each sample X^i, $i = 1, \ldots, N$, another sample $\tilde{X}^i$ that has the same distribution as X^i, but is negatively correlated with X^i. This has two effects: first of all, we increase the number of samples used for importance sampling and second, as the new samples are negatively correlated with the old samples, the Monte-Carlo variance is reduced. The computation of these samples is usually much faster than creating new samples, which requires the use of the expensive FFBS or simulation smoother algorithms.

Definition 3.6 (antithetic variable)**.** Let $X, \tilde{X} \in \mathbf{R}^k$ be two random variables with the same distribution, $\mathcal{L}(X) = \mathcal{L}(\tilde{X})$ and $f : \mathbf{R}^k \to \mathbf{R}$. Then $\tilde{X}$ is called an antithetic variable of X for f, if $\operatorname{Cov}\Big(f(\tilde{X}), f(X)\Big) < 0$. If $k = 1$ and f is the identity, we just say that $\tilde{X}$ is an antithetic variable of X.

(Durbin and Koopman, 1997) introduce two antithetic variables: balanced for location and balanced for scale, both of which are tailored to the multivariate normal distribution.

Definition 3.7 (antithetic variable balanced for location and scale, (Durbin and Koopman, 1997))**.** Let $X \sim \mathcal{N}(\mu, \Sigma)$ for $\mu \in \mathbf{R}^k$ and $\Sigma \in \mathbf{R}^{k\times k}$ positive definite. We call

$$\tilde{X} = \mu + (\mu - X) \tag{3.40}$$

the antithetic balanced for location. If $L \in \mathbf{R}^{k\times k}$ is a Cholesky root of Σ and

$$X = \mu + L\varepsilon$$

with $\varepsilon \sim \mathcal{N}(0, I)$, let $c = \varepsilon^T\varepsilon \sim \chi^2_k$ and $c' = F^{-1}_{\chi^2_k}(1 - F_{\chi^2_k}(\sqrt{c}))$. We call

$$\check{X} = \mu + \sqrt{\frac{c'}{c}}\,(X - \mu) \tag{3.41}$$

the antithetic balanced for location.

Lemma 3.12. *In the above definition, $\tilde{X}_i$ is an antithetic variable of X for the coordinate functions $f_i : \mathbf{R}^k \to \mathbf{R}$, $f_i(x) = x_i$, $i = 1, \ldots, k$. Furthermore, $\tilde{c}$ is an antithetic variable of c.*

Proof. It is easy to see that $\tilde{X}$ has the same distribution as X. Furthermore

$$\operatorname{Cov}\Big(f_i(X), f_i(\tilde{X})\Big) = \operatorname{Cov}(2\mu_i - X_i, X_i) = -\Sigma_{i,i} < 0.$$

For c and $\tilde{c}$, let $U = F_{\chi^2_k}(c)$, then $U \sim \operatorname{Unif}(0,1)$ and $\tilde{U} = 1 - U = F_{\chi^2_k}(\tilde{c})$. As $\tilde{U} \sim \operatorname{Unif}(0,1)$ as well, $\mathcal{L}(c) = \mathcal{L}(\tilde{c})$. In (Whitt, 1976, Lemma 2.3) it is shown that for any pair of real-valued random variables (Y, W) with CDF H and marginal CDFs F, G, it holds

$$\operatorname{Cov}(Y, W) = \int_{\mathbf{R}^2} H(y, w) - F(y)G(w)\,\mathrm{d}y\,\mathrm{d}w,$$

and, furthermore, by (Whitt, 1976, Theorem 2.1 and Lemma 2.4) that the joint CDF of $(c, \tilde{c})$ is $(y, w) \mapsto \max\{0, F(y) + G(w) - 1\}$, where F is the CDF of c and G the CDF of $\tilde{c}$. As

$$a + b - 1 = ab + a(1 - b) + b - 1 = ab - (1 - a)(1 - b) < ab$$

for all $a, b \in (0,1)$, we have

$$\begin{aligned}\operatorname{Cov}(c, \tilde{c}) &= \int_{\mathbf{R}^2} H(y,w) - F(y)G(w)\,\mathrm{d}y\,\mathrm{d}w \\ &= \int_{\mathbf{R}^2} \max\{0, F(y)+G(w)-1\} - F(y)G(w)\,\mathrm{d}y\,\mathrm{d}w < 0.\end{aligned}$$

□

Let us mention that, by the properties of the standard multivariate normal distribution, $c = \|u\|$ and $\frac{u}{\|u\|}$ are independent. Writing

$$X = \mu + \|u\| L \frac{u}{\|u\|} = \mu + \|u\| \frac{X-\mu}{\sqrt{c}},$$

we see that

$$\breve{X} = \mu + \sqrt{\tilde{c}} \frac{X-\mu}{\sqrt{c}}$$

has the same distribution as X, as $\tilde{c} \sim \mathcal{L}(\|u\|^2)$ and is independent of $\frac{X-\mu}{\sqrt{c}}$.

Given a GLSSM-proposal and samples $X^1, \ldots, X^N$ from it, we can cheaply calculate these antithetic variables: for the location balanced antithetic we can calculate the mean using the Kalman-smoother and for the scale balanced antithetic we can calculate c and c' using the inverse CDF of the χ^2_k distribution and the standard normal samples used to sample X^i in the first place, for which fast implementations are readily available. Incidental, we obtain a third antithetic,

$$\breve{X} = \mu - \sqrt{\frac{c'}{c}}(X-\mu) \tag{3.42}$$

for free. We can then estimate the log-likelihood in Equation (3.39) by replacing each occurrence of $\tilde{w}_\theta(X^i)$ by

$$\frac{1}{4}\left(\tilde{w}_\theta(X^i) + \tilde{w}_\theta(\tilde{X}^i) + \tilde{w}_\theta(\breve{X}^i) + \tilde{w}_\theta(\breve{X}^i)\right). \tag{3.43}$$

As the procedure to evaluate the likelihood by importance sampling becomes expensive as the dimension of the model increases, (Durbin and Koopman, 1997) recommend finding an initial value $\hat{\theta}_0$ by maximizing a deterministic version of Equation (3.36). For this, denote by s^* the mode of the linear signal, conditional on the pseudo-observations z. As S follows a multivariate Gaussian, s^* is also the mean which can be computed efficiently by the Kalman or signal-smoother. Approximating the conditional expectation in Equation (3.37) by $w_\theta(s^*)$ then yields

$$\log p_\theta(y) \approx \log g_\theta(z) + \log w_\theta(s^*), \tag{3.44}$$

which can be evaluated without simulation by the LA. A better approximation can be obtained by performing a fourth-order Taylor expansion of $s \mapsto w_\theta(s)$ around the mode s^*, which yields

$$\log p_\theta(y) \approx \log g_\theta(z) + \log w_\theta(s^*) + \log\left(1 + \frac{1}{8}\sum_{t=1}^{n}\sum_{j=1}^{m} l^{(4)}_{t,j}(s^*) v^2_{t,j}\right), \tag{3.45}$$

where $l^{(4)}$ is the fourth derivative of the log-weights $s \mapsto \log w_\theta(s)$ and $v_{t,j}$ is the conditional variance $\operatorname{Var}(S_{t,j}|Z=z)$ in the proposal. Again, we refer the interested reader to the source for the details.

The resulting procedure to find the MLE $\hat{\theta}$ in a PGSSM with linear signal is summarized in Section 3.7. Notice that we use CRNs to ensure numerical convergence. The numerical optimization can be performed using any standard solver such as the BFGS algorithm (Nocedal and Wright, 2006, Chapter 6.1). We cannot give guarantees that this procedure produces the true MLE, i.e. finds

Algorithm 9 Maximum likelihood estimation in a PGSSM with linear signal using EIS.

Require: parameterized PGSSM with linear signal, initial $\theta^0 \in \Theta$, observations $y \in \mathbf{R}^{(n+1)p}$, number of samples N

function APPROX_LOGLIK(θ)
  obtain LA of the PGSSM for θ ▷ Algorithm 5
  obtain mode s^* and conditional variances $v_{t,j}$ from the LA ▷ Algorithms 1 and 2
  return approximate log-likelihood ▷ Equation (3.44) or Equation (3.45)
end function

function ESTIMATE_LOGLIK(θ)
  obtain LA of the PGSSM for θ ▷ Algorithm 5
  obtain EIS proposal $\mathbf{G}_{(z,\Omega)}$ ▷ Algorithm 6, LA as initial values
  sample N signals S^i from $S|Z = z$ in EIS ▷ Algorithm 3 or signal smoother
  obtain mode s^* in EIS proposal ▷ Algorithm 2 or signal smoother
  calculate antithetic variables $\tilde{S}^i, \check{S}^i, \breve{S}^i$ ▷ Equations (3.40) to (3.42)
  set $\tilde{w}_\theta^i = \frac{1}{4}\left(\tilde{w}_\theta(X^i) + \tilde{w}_\theta(\tilde{X}^i) + \tilde{w}_\theta(\check{X}^i) + \tilde{w}_\theta(\breve{X}^i)\right)$ Equation (3.43)
  set $\tilde{w}_\cdot = \frac{1}{N}\sum_{i=1}^N \tilde{w}_\theta^i$
  set $\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^N \left(\tilde{w}_\theta^i - \tilde{w}_\cdot\right)^2$
  calculate $\widehat{\log g_\theta(z)}$ ▷ Algorithm 1
  return $\widehat{\log p_\theta(y)}$ ▷ Equation (3.39)
end function

maximize APPROX_LOGLIK with initial value θ^0 ▷ numerically
set θ^0 to optimal value
maximize ESTIMATE_LOGLIK with initial value θ^0 and CRNs ▷ numerically
set $\hat{\theta}$ to optimal value
return $\hat{\theta}$

the global maximizer. However, as we have discussed earlier, we are not interested in frequentist properties of $\hat{\theta}$ but see the estimation procedure as a hyperparameter tuning step. Thus, a local maximum may well be sufficient. Nevertheless, checking different starting points and random number seeds should be used to get as close as possible to the global maximum.

Notice that our discussion implies that we cannot reuse a GLSSM proposal used for θ at another θ', as $p_{\theta'}(x) \neq g_\theta(x)$. While we can still calculate the weights using the general Equation (3.36), we presume that the old proposal is not a good choice for the new target. The reason for this is that θ will usually contain parameters related to the covariance structure of the innovations and observations, and these parameters usually affect many, if not all states or observations. For example, it is common to model states that perform a random walk with common innovation variance σ^2 as an element of θ. As the distributions lie in a high-dimensional space, slight misspecification of the covariance structure will drastically deteriorate the performance of importance sampling.

If computations are so involved that we want to avoid running the optimal importance sampling scheme as much as possible, one could try, if the model under investigation allows for it, to split θ into (θ_x, θ_y) where θ_x only affects the state transitions and θ_y only affects the observation densities. Then a coordinate ascent scheme could be employed, where the update step for θ_y can reuse the proposal, provided that θ_y does not change too much and the observation density $p_\theta(y|x)$ is not too sensitive to changes in θ_y, which should imply that the proposal is still close enough to give good importance sampling performance. Then numerical differentiation is only required to update θ_x.

3.8 Comparison of Importance Sampling method

We now have three tools to produce Gaussian importance sampling proposals: the LA, the CE-method and EIS. Naturally, we want to choose the optimal tool for the problem at hand. In this section, we investigate under which circumstances which method is to be preferred over the others. To judge the performance of each method, we will discuss the following quality criteria:

- breakdown of methods,
- time and space complexity of the method,
- speed of stochastic convergence, as indicated by the asymptotic variance, for the CE-method and EIS,
- speed of numerical convergence, as indicated by the number of iterations until Algorithms 6 and 8 reach numerical convergence for fixed sample size N and precision ϵ, and
- performance of the optimal proposal, as measured by the efficiency factor, especially as n or m comes larger.

Let us elaborate on these criteria. With a breakdown of the methods, we mean settings in which either the numerical scheme diverges, produces parameters that lead to invalid proposals, i.e. negative variances, or where the proposals fail to produce consistent importance sampling estimates. Time and space complexity allow us to compare the methods theoretically, i.e. be independent of implementation details. The speed of stochastic convergence is relevant as well: The smaller the asymptotic variance, the smaller we can choose the sample size N and thus decrease computation time. Similarly, numerical convergence directly affects computation time.

reformulate this paragraph nicer

Finally, if one method has vastly better performance at the optimum, we might be willing to spend more time initially to save time later when we use the proposal to perform inference. Of special interest is the performance for long (large n) or fat (large m) time series, as the models we fit in Chapter 4 usually fall into one of these categories.

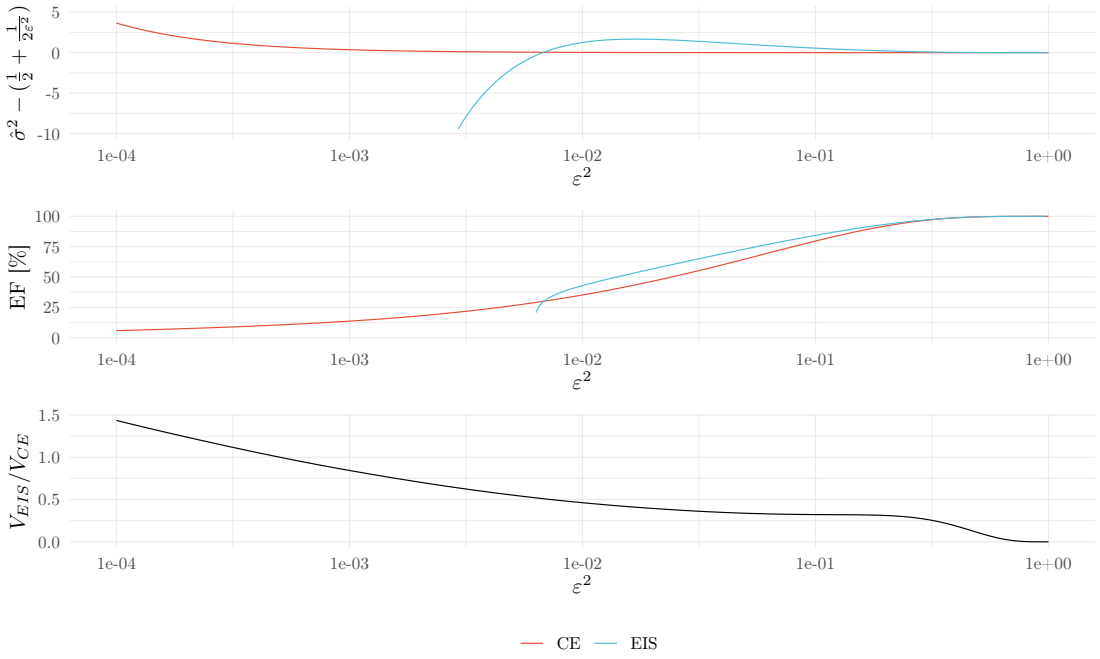

Figure 3.3: TODO

3.8.1 Breakdown of methods

Let us start with a classical example in which the LA fails to produce consistent importance sampling estimates.

Example 3.3 (Failure of LA)**.** Consider the Gaussian scale mixture $\mathbf{P} = \frac{1}{2}\left(\mathcal{N}(0,1) + \mathcal{N}(0,\varepsilon^{-2})\right)$ with mode $x^* = 0$, this is the same setup as in Example 3.2. The LA is $\mathbf{G}_{\text{LA}} = \mathcal{N}\left(0, \frac{1}{\varepsilon^2 - \varepsilon + 1}\right)$, whose variance goes to 1 as ε goes to 0, so the LA will miss close to $\frac{1}{2}$ of the total mass. For ε small enough, the variance of the LA will be smaller than $\frac{1}{2\varepsilon^2}$, whence the second moment of the weights is infinite and importance sampling with $\mathbf{G}_{\text{LA}}$ is inconsistent.

The CE-method minimizes the KL-divergence between $\mathbf{P}$ and $\mathbf{G}_\psi$, is given by $\mathbf{G}_{\text{CE}} = \mathcal{N}(0, \sigma^2)$, where $\sigma^2 = \frac{1}{2}\left(1 + \varepsilon^{-2}\right)$ is the variance of $\mathbf{P}$. As $\sigma^2 > \frac{1}{2}\varepsilon^{-2}$, the weights have finite second moment, and importance sampling with $\mathbf{G}_{\text{CE}}$ is consistent.

add proof for $\frac{1}{2}$ to appendix

As EIS does not yield analytically tractable proposals in this setting, we resort to a simulation study. Using the same setup as described in Example 3.4, we replicate $M = 100$ times $\hat{\psi}_{\text{CE}}$ and $\hat{\psi}_{\text{EIS}}$ for varying levels of ε^2. The resulting excess variances, i.e. $\sigma^2 - \left(\frac{1}{2} + \frac{1}{2\varepsilon^2}\right)$, efficiency factors and asymptotic efficiencies are displayed in Figure 3.3. We see that for small ε^2, EIS is inconsistent, while the CE-method stays consistent. However, as is to be expected, for small ε^2, the efficiency factor becomes very small.

This is more of a technical counter-example, in practice the LA produces good importance sampling proposals, especially for LCSSMs.

In the LCSSM setting EIS may produce invalid proposals, as estimates of the variance component in the weighted least squares regression are not guaranteed to be negative. Thus EIS may produce negative variances. To deal with this, the original EIS paper Richard and Zhang, 2007, Section 3.2 recommends either inflating the prior or setting the parameters in question to arbitrary fixed values. Alternatively using a more expensive constrained linear least squares solver, such as a conjugate-gradient method Branch, Coleman, and Y. Li, 1999 or the BVLS (bounded variable least

method	single iteration (time)	single iteration (space)	simulation (time)
LA	$\mathcal{O}(n\,p^3)$	$\mathcal{O}(np^2)$	$\mathcal{O}(n(p^3+m^3+N\,m^2))$
EIS	$\mathcal{O}(n(m^2+p^3+N\,p^2))$	$\mathcal{O}(N\,p+n(p^2+m^2))$	$\mathcal{O}(n(p^3+m^3+N\,m^2))$
CE-method	$\mathcal{O}(n(Nm^2+m^3))$	$\mathcal{O}(Nm+nm^2)$	$\mathcal{O}(N\,n\,m^2)$

Table 3.2: Computational complexities of importance sampling algorithms.

squares) solver Stark and Parker, 1995 may be appropriate, as is re-running the EIS procedure with a different random seed. Finally, in the LCSSM setting, we could also identify the corresponding observation as missing, similar to the argument presented in Section 3.6.1 for the CE-method.

The CE-method presented in Section 3.6.2 (Algorithm 8) depends on the fact that the covariance matrix of the posterior $\mathrm{Cov}\,(X|Y=y)$ is symmetric positive definite (SPD), i.e. non-singular. This might be violated if, e.g., the model contains seasonal components whose associated innovations have variance 0. In this case, the Cholesky roots involved will not be unique. Still Algorithm 8 will, as $N \to \infty$ converge a globally optimal solution, though it may not be unique.

3.8.2 Computational complexity

Throughout this section, we assume that the model in question is a LCSSM with linear signal (c.f. Definition 3.5) to simplify the treatment. This benefits the LA and EIS approaches, as they may then be implemented in terms of the simulation and signal smoother. If the observation dimension p is smaller than that of states m, this is more efficient and we'll assume this as well. An overview of computational complexities is given in Table 3.2. Note that most operations can be parallelized in one way or the other, e.g. sampling from the proposals, and so the time-complexities are not necessarily indicative of real-world-performance. Still they provide theoretical insight into the performance of the three methods considered.

Let us begin with a discussion of the computational complexity involved in finding the optimal parameters, $\psi_{\mathrm{LA}}, \hat{\psi}_{\mathrm{EIS}}$ and $\hat{\psi}_{\mathrm{CE}}$. Here we focus on a single iteration and treat the number of iterations empirically in Section 3.8.4.

As the LA is based on the Kalman-smoother, the time complexity of a single iteration is $\mathcal{O}(n(m^2+p^3))$. The CE-method and EIS need to generate N samples from the current proposal. For the CE-method this amounts to $\mathcal{O}(N\,n\,m^2)$ operations (see Section 3.6.2). For EIS, using the simulation smoother Durbin and Koopman, 2002 requires $\mathcal{O}(n(m^2+p^3+N\,p^2))$ operations: we need to run the Kalman filter once, while preparing the matrices required for the simulation smoother. Then, provided Cholesky roots of the innovation covariance matrices Σ_t are already available, only matrix-vector multiplications are necessary for the simulation smoother. Obtaining the EIS model parameters is efficient, requiring only $\mathcal{O}(n(N\,p^2+p^3))$ operations for constructing the n $p\times p$ design matrices and estimating the optimal parameters.

Another concern is the time required to generate N samples from the fitted model. For both the LA and EIS this requires using either the simulation smoother or the FFBS algorithm. This necessitates inverting $p\times p$ matrices in the Kalman filter and $m\times m$ matrices when simulating the states. Fortunately, these steps can be performed offline, after which the simulation of a single sample requires only $\mathcal{O}(n)$ matrix-vector multiplications. The CE-method simulation is based on applying Equation (3.25), which only requires $\mathcal{O}(n\,m^2)$ time per sample.

Concerning space complexity, the LA has to run the Kalman filter with $\mathcal{O}(n(p^2+m^2))$ space and store $\mathcal{O}(np)$ parameters. EIS has the same space requirement, but needs additional $\mathcal{O}(Np)$ storage for the simulated signals. As the weights w_t in EIS depend only on the current signals $S_t^1,\ldots,S_t^N$, they may be discarded afterwards. See Section 3.6.2 for the derivation of the $\mathcal{O}(Nm+nm^2)$ space requirement of the CE-method.

The LA has the fastest and most space-efficient iteration of the three methods because it does not require the simulation of N samples. This makes it an ideal candidate as an initial guess for the other two methods. For $p \ll m$, EIS is faster than CE-method as it is based on the signals

S only, thus having access to the efficient simulation and signal smoother algorithms. The same is true for the space complexity. If, however, $p \approx m$, there is no linear signal or the observations are not conditionally independent given the states or signals, the speed of EIS and CE-method should be comparable. While theoretically, the CE-method performs sampling faster than the other two methods, for large numbers of samples N the difference is negligible because the additional computations only have to be performed once.

3.8.3 Asymptotic variance

As we have seen in the previous section, the number of samples N used to estimate ψ_{CE} and ψ_{EIS} enter linearly into the computational complexities. Naturally, we want to know how big a sample size we should choose for our procedures and whether one of the two simulation-based procedures requires fewer samples than the other. To answer this question we turn to the two central limit theorems, Theorems 3.6 and 3.9. If N is large, the asymptotic variances (or rather: the asymptotic standard deviations) tell us how much stochastic variation we should expect around the optimal value, and can thus guide us in choosing N. We start with two examples in a univariate setting, where both the CE-method and EIS use Gaussian proposals with either fixed variance (Example 3.4) or mean (Example 3.5). This allows us to compare the methods for either the mean (variance) if the variance (mean) is fixed and potentially misspecified, i.e. not the global optimum. Additionally, the univariate setting allows us, in some cases, to derive analytical expressions of the efficiencies involved, allowing us to interpret them.

rewrite this more clearly

To compare both methods we will determine the asymptotic relative efficiencies, i.e. $\frac{\mathrm{Var}\left(\hat{\psi}_{\mathrm{EIS}}\right)}{\mathrm{Var}\left(\hat{\psi}_{\mathrm{CE}}\right)}$, with values smaller than 1 indicating that EIS requires (asymptotically) fewer samples for the same precision as the CE-method. Let us note that we are comparing the efficiencies of parameters ψ, not those of derived parameters such as the standard deviation or the ESS. However, should both methods have the same optimal value the relative efficiencies are the same for all parameters derived from ψ, by the delta method. By a continuity argument, the same is approximately true if the optimal values of the CE-method and EIS are close.

Example 3.4 (univariate Gaussian, σ^2 fixed)**.** Consider the probability space $(\mathbf{R}, \mathcal{B}(\mathbf{R}), \mathbf{P})$ where $\mathbf{P} = p\lambda$ for the Lebesgue measure λ which is symmetric around 0, i.e. $p(-x) = p(x)$ for λ-a.e. $x \in \mathbf{R}$ and possesses up to third order moments. Let $\mathbf{G} = \mathbf{P}$, so $W \equiv 1$ and let $\mathbf{G}_\psi = \mathcal{N}\left(\sigma\psi, \sigma^2\right)$ be the single parameter natural exponential family of Gaussians with fixed variance $\sigma^2 > 0$. Then

$$\log g_\psi(x) = \psi T(x) - \frac{\psi^2}{2} + \log h(x),$$

where $T(x) = \frac{x}{\sigma}$ and $h(x)$ is the density of $\mathcal{N}(0, \sigma^2)$ w.r.t. Lebesgue measure. Note that T is centered under $\mathbf{P}$. To compare the asymptotic behavior of the CE-method and EIS we compute the asymptotic variances arising from their respective central limit theorems (Theorems 3.6 and 3.9).

By symmetry, both ψ_{CE} and ψ_{EIS} are equal to 0. Then $I(\psi) = 1$ for all ψ, so

$$V_{\mathrm{CE}} = \mathrm{Cov}_{\mathbf{P}}(T) = \frac{\tau^2}{\sigma^2}, \tag{3.46}$$

where $\tau^2 = \mathbf{P}\,\mathrm{id}^2$ is the second moment of $\mathbf{P}$.

Additionally, $B_{\mathrm{EIS}} = (\mathrm{Cov}_{\mathbf{P}}(T))^{-1} = \frac{\sigma^2}{\tau^2}$ and

$$\begin{aligned} M_{\mathrm{EIS}} &= \mathrm{Cov}_{\mathbf{P}}\left((\log\frac{p(x)}{h(x)} - \lambda_{\mathrm{EIS}})T\right) \\ &= \mathrm{Cov}_{\mathbf{P}}\left((\log p - \log h - \mathbf{P}(\log p - \log h))\,T\right) \\ &= \frac{1}{\sigma^2}\int p(x)x^2\left(\log p(x) + \frac{x^2}{2\sigma^2} - \mathbf{P}\left(\log p(x) + \frac{\tau^2}{2\sigma^2}\right)\right)^2 \mathrm{d}x. \end{aligned}$$

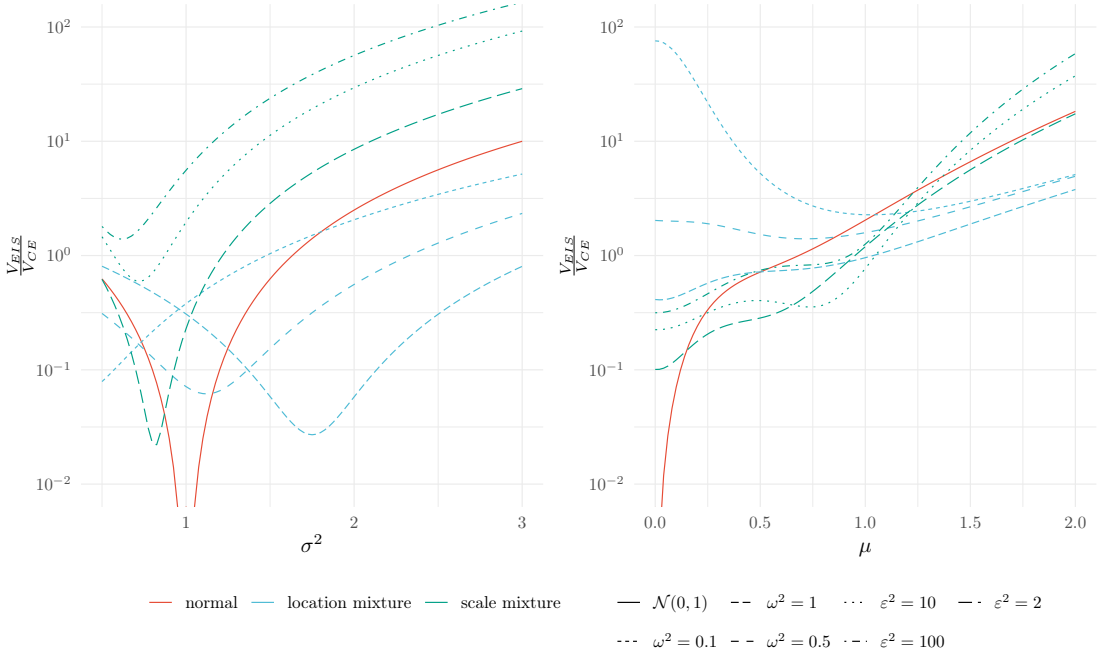

Figure 3.4: Asymptotic relative efficiency $\frac{V_{\text{EIS}}}{V_{\text{CE}}}$ for the normal distribution from Example 3.4 (left hand side) and Example 3.5 (right hand side). Here $\mathbf{P}$ is either the standard normal distribution, a Gaussian location mixture, or a Gaussian scale mixture. $\mathbf{G}_\psi$ is the normal distribution $\mathcal{N}(\mu, \sigma^2)$, where either σ^2 is fixed (left) and μ determined by the CE-method / EIS, or the other way around (right). Notice the log scale of the y-axis. As μ or σ^2 get close to their true values, EIS outperforms the CE-method in terms of asymptotic variance, see Proposition 3.5. todo: clean up figure legend / linetype, order eps and omega, add global σ^2 choosen by estimating both parameters at the same time?

Thus

$$V_{\text{EIS}} = B_{\text{EIS}} M_{\text{EIS}} B_{\text{EIS}} = \sigma^2 \frac{\gamma}{\tau^4},$$

where $\gamma = \int p(x) x^2 \left(\log p(x) + \frac{x^2}{2\sigma^2} - \mathbf{P}(\log p(x) + \frac{\tau^2}{2\sigma^2}) \right)^2 \mathrm{d}x$.

Let us now consider three exemplary choices of $\mathbf{P}$ that illustrate a target that is sufficiently well-behaved (the standard normal), multimodal (a Gaussian location mixture) and has different behavior in the tails than indicated at the mode (a Gaussian scale mixture). For each target, we vary σ^2 from $\frac{1}{2}$ to 3 and obtain relative efficiencies of the CE-method and EIS either analytically or by simulation, the results are shown in the left-hand side of Figure 3.4.

Normal distribution If $\mathbf{P} = \mathcal{N}(0, \tau^2)$ is a normal distribution, this reduces to

$$V_{\text{EIS}} = \frac{5}{2} \left(\frac{\tau^2}{\sigma^2} - 1 \right)^2 \frac{\sigma^2}{\tau^2} = \frac{5}{2} \frac{(V_{\text{CE}} - 1)^2}{V_{\text{CE}}}$$

and so for $\tau^2 = \sigma^2$ $\hat{\psi}_{\text{EIS}}$ converges faster than the standard $\mathcal{O}(N^{-\frac{1}{2}})$ rate. Indeed in this case $\hat{\psi}_{\text{EIS}} = \psi_{\text{EIS}}$ a.s. for $N > 1$, see Proposition 3.5.

Gaussian location mixture Consider now the case where $\mathbf{P} = \frac{1}{2}\mathcal{N}(-1, \omega^2) + \frac{1}{2}\mathcal{N}(1, \omega^2)$ is a Gaussian location mixture. The second moment is $\tau^2 = 1 + \omega^2 = -\frac{1}{2\psi_{\text{CE}}}$. Unfortunately, there is no closed-form expression for many of the terms required for the analysis EIS. Instead, we resort to a simulation study to determine the asymptotic variances and relative efficiencies for three different values of $\omega^2 \in \{0.1, 0.5, 1.0\}$.

To this end we draw $M = 100$ times from the distribution of $\hat{\psi}_{\text{CE}}$ and $\hat{\psi}_{\text{EIS}}$, where we use $N = 1000$ samples from the tractable $\mathbf{P}$ as importance samples. We only iterate a single time for both procedures. From individual estimates, we estimate the asymptotic variances V_{CE} and V_{EIS} by the respective empirical variances, and determine the relative efficiency of EIS over the CE-method as $\frac{V_{\text{EIS}}}{V_{\text{CE}}}$. Again, we vary the fixed variance of the proposals, σ^2, from $\frac{1}{2}$ to 3.

discuss MC error of this estimate, small enough to ignore?

Gaussian scale mixture Finally we consider $\mathbf{P} = \frac{1}{2}\left(\mathcal{N}(0,1) + \mathcal{N}(0,\varepsilon^{-2})\right)$ for $\varepsilon^2 \in \{2, 10, 100\}$, a scale mixture similar to the one seen in Example 3.3. Contrary to that example, we choose ε big, making the $\mathcal{N}(0,1)$ component the one with large variance, to make importance sampling with σ^2 in the range considered consistent. Here $\tau^2 = \frac{1}{2} + \frac{1}{2\varepsilon^2}$. Again, we estimate the asymptotic V_{EIS} in the same way as for the Gaussian location mixture, with $M = 100$ estimates using $N = 1000$ samples each.

Note that for fixed σ^2 the asymptotic variance of the CE-method V_{CE} is the same in all of the examples considered, as we sample directly from the tractable $\mathbf{P}$, so V_{CE} only depends on $\mathbf{P}$ through its second moment τ^2. The asymptotic variance of EIS however depends on both τ^2, as well as γ, which depends on global properties of $\mathbf{P}$.

From the left-hand side of Figure 3.4 we can observe that in the case of $\mathbf{P} = \mathcal{N}(0,1)$ EIS has smaller asymptotic variance compared to the CE-method, as long as σ^2 is not heavily misspecified. Indeed, if $\sigma^2 = 1$ is correctly specified, by Proposition 3.5, EIS has asymptotic variance 0 and converges already for a single sample.

Consider now the case where $\mathbf{P}$ is a Gaussian location mixture. For $\omega^2 = 1$, the location mixture is unimodal with variance 2 and EIS outperforms the CE-method in terms of asymptotic variance in the range considered. For the smaller values of ω^2 considered here, the location mixture is bimodal. Close to the true variance $1 + \omega^2$, EIS still outperforms the CE-method.

For the Gaussian scale mixture, the case is less clear. Here the true variance is $\frac{1}{2} + \frac{1}{2\varepsilon^2}$. The location of the minimal relative efficiency is still close to this true variance, however, as ε^2 grows, the CE-method starts to dominate EIS. Additionally, recall from Example 3.3 that for large ε^2 EIS becomes inadmissible.

Example 3.5 (univariate Gaussian, μ fixed)**.** Consider the same setup as in Example 3.4, i.e. $\mathbf{P}$ is symmetric around 0 with second moment τ^2, but let $\mathbf{G}_\psi = \mathcal{N}(\mu, -\frac{1}{2\psi})$ be the single parameter natural exponential family of Gaussians with fixed mean μ and variance $\sigma^2 = -\frac{1}{2\psi}$.

Then

$$\log g_\psi(x) = \psi T(x) + \frac{1}{2}\log(-2\psi) - \frac{1}{2}\log 2\pi$$

for $T(x) = (x-\mu)^2$. Thus $\mathbf{P}T = \tau^2 + \mu^2$ and $\operatorname{Cov}_{\mathbf{P}} T = \nu - \tau^4 + 4\tau^2\mu^2$ where $\nu = \mathbf{P}\,\mathrm{id}^4$ and $\tau^2 = \mathbf{P}\,\mathrm{id}^2$.

By matching moments, we obtain $\psi_{\text{CE}} = -\frac{1}{2(\tau^2+\mu^2)}$ and $I(\psi_{\text{CE}}) = \frac{1}{2\psi_{\text{CE}}{}^2} = 2(\tau^2+\mu^2)^2$. In total

$$V_{\text{CE}} = \frac{1}{4(\tau^2+\mu^2)^4}\left(\nu - \tau^4 + 4\tau^2\mu^2\right) \tag{3.47}$$

For EIS,

$$\begin{aligned}\psi_{\text{EIS}} &= \left(\operatorname{Cov}_{\mathbf{P}} T\right)^{-1} \operatorname{Cov}_{\mathbf{P}}(T, \log p) \\ &= \left(\nu - \tau^4 + 4\tau^2\mu^2\right)^{-1} \underbrace{\int p(x)((x-\mu)^2 - \tau^2 - \mu^2)(\log p(x) - \mathbf{P}\log p(x))\,\mathrm{d}x}_{=\gamma}.\end{aligned}$$

Then

$$V_{\text{EIS}} = \left(\nu - \tau^4 + 4\tau^2\mu^2\right)^{-2} \mathbf{P}\left((\mathrm{id}-\mu)^4\left(\log p - \psi_{\text{EIS}}(\mathrm{id}-\mu)^2 - \mathbf{P}\log p + \psi(\tau^2+\mu^2)\right)^2\right).$$

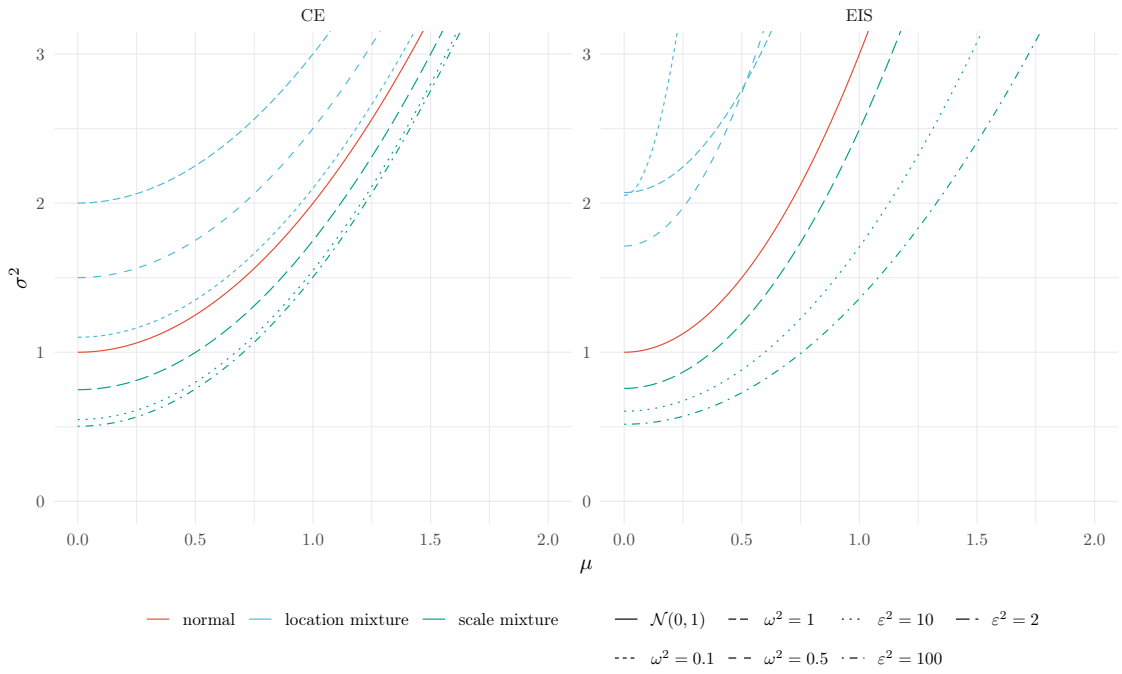

Figure 3.5: TODO

We now perform the same analysis as in Example 3.4, the resulting ratio of asymptotic variances is displayed in the right-hand side of Figure 3.4. In general, the variances $\sigma^2_{\text{CE}} = -\frac{1}{2\psi_{\text{CE}}}$ and $\sigma^2_{\text{EIS}} = -\frac{1}{2\psi_{\text{EIS}}}$ are different, so the ratio is no longer an asymptotic relative efficiency. However, it is still relevant as a measure of the relative speed of stochastic convergence of both methods. Additionally, we display the resulting optimal variances in Figure 3.5.

Normal distribution For the normal distribution $\mathbf{P} = \mathcal{N}(0, \tau^2)$ where $\nu = 3\tau^4$ and $\gamma = -\tau^2$, so

$$\psi_{\text{EIS}} = \frac{-\tau^2}{2\tau^2\left(\tau^2 + 2\mu^2\right)} = \frac{-1}{2(\tau^2 + 2\mu^2)}.$$

Thus the EIS proposal uses variance $\sigma^2_{\text{EIS}} = \tau^2 + 2\mu^2$, which is bigger than the variance of $\sigma^2_{\text{CE}} = \tau^2 + \mu^2$ optimal for the CE-method.

In this case the asymptotic variances are

$$V_{\text{CE}} = \frac{\tau^2(\tau^2 + 2\mu^2)}{2\left(\tau^2 + \mu^2\right)^4}$$

and

$$V_{\text{EIS}} = \frac{\mu^2\left(2\mu^6 + 45\mu^4\tau^2 + 15\tau^6\right)}{4\tau^4\left(2\mu^2 + \tau^2\right)^4},$$

see the Appendix for details.

reference it

Gaussian location mixture same setup as before

Gaussian scale mixture same setup as before

On the left-hand side of Figure 3.4 we see that for μ close to the optimal value, EIS has smaller asymptotic variance than the CE-method, except for the two bimodal location measures. Again, due to the finite sample convergence of EIS, Proposition 3.5, the asymptotic variance V_{EIS} goes to 0 as $\mu \to 0$. The more μ becomes misspecified, the ratio of asymptotic variances starts to grow.

In Figure 3.5 we see that, except for the extreme scale mixtures, EIS tends to produce proposals that have a larger variance than those produced by the CE-method. As we will see in the discussion of Figure 3.7, this might be advantageous for EIS as proposals with a small variance run the risk of missing a large part of the probability mass of the target.

clean this

In applications, e.g. the model studied in Chapter 4, we are interested in the performance of the importance sampling proposals generated by the LA, CE-method and EIS under more complex circumstances than those discussed in Examples 3.4 and 3.5. In particular, the dimension of ψ is high ($\mathcal{O}(n \cdot m)$ or even $\mathcal{O}(n \cdot m^2)$) and proposals may not come from a natural exponential family, so analysis based on Theorems 3.6 and 3.9 is not possible.

really?

Instead, we resort to simulation studies to gain insights into the circumstances when one should prefer one method over the other. As a leading example, we will use the following vector-autoregressive state space model with negative binomial observations. A similar, though more involved, model is studied in Section 4.2 with real data.

Example 3.6 (Negative Binomial VAR(1) SSM)**.** In this example, we consider a SSM where states X_t follow a stationary Gaussian VAR(1) process, initialized in its stationary distribution $\mathcal{N}(0, \Sigma)$ for SPD Σ. For simplicity let the transition matrices be given by a multiple of the identity, i.e. $A_t = \alpha I_m$ for all t where $\alpha \in (-1, 1)$

add I to symbols

. In total, the states are governed by

$$\begin{aligned} X_0 &\sim \mathcal{N}(0, \Sigma) \\ X_t &= \alpha X_{t-1} + \varepsilon_t \\ \varepsilon_t &\overset{\text{i.i.d}}{\sim} \mathcal{N}(0, (1-\alpha^2)\Sigma), t = 1, \dots, n \end{aligned}$$

where the $\varepsilon_1, \dots, n$ and X_0 are jointly independent. The observations follow a conditional negative binomial distribution

$$Y_t^i | X_t \sim \text{NegBinom}\left(\exp(X_t^i), r\right), \qquad i = 1, \dots, p \qquad t = 0, \dots, n$$

and individual observations are conditionally independent given the current state. The parametrization of the negative binomial distribution $\text{NegBinom}(\mu, r)$ is such that the density is

$$p_{\mu,r}(y) = \binom{y+r-1}{r} \left(\frac{\mu}{r+\mu}\right)^y \left(\frac{r}{r+\mu}\right)^r \propto \mu^y (\mu + r)^{-(r+y)},$$

where proportionality is in μ, with expectation μ, variance $\mu + \frac{\mu^2}{r}$ and support $\mathbf{N}_0$.

Our first simulation study concerns the non-asymptotic behavior of the CE-method and EIS estimators, i.e. finite sample analogs of Theorems 3.6 and 3.9. To this end, we let $m = 1$ in Example 3.6 and fix n to

...

. We then simulate once from the marginal distribution of Y and perform the LA to a prespecified precision ϵ and maximum number of iterations n_{iter}, obtaining a proposal distribution $\mathbf{G}_{\text{LA}}$. Using

a large number of samples N_{true} from this proposal we find the optimal $\mathbf{G}_{\text{CE}}$ and $\mathbf{G}_{\text{EIS}}$ using the same desired precision and number of iterations as for the LA. For the remainder of this section, we ignore sampling variation in these proposals and treat them as exact.

To determine the non-asymptotic sampling behavior we now perform the above procedure again, using only $N \ll N_{\text{true}}$ many samples for both procedures, obtaining proposals $\hat{\mathbf{P}}^N_{\text{CE}}$ and $\hat{\mathbf{P}}^N_{\text{EIS}}$. As the full proposals are Gaussian distributions on $\mathbf{R}^{(n+1)\times m}$, either given as the posterior of a GLSSM (LA, EIS) or by a Gaussian Markov process(CE-method), see Section 3.6. This procedure is repeated M times for every sample size N considered, with different initial random seeds, obtaining $\hat{\mathbf{P}}^{N,i}_{\text{CE}}$ and $\hat{\mathbf{P}}^{N,i}_{\text{EIS}}$ for $i = 1, \dots, M$.

To assess the speed of convergence of the CE-method and EIS we then estimate the mean squared error of means and variances of the $(n+1) \times m$ univariate marginals as N, the number of samples used to obtain $\hat{\psi}_{\text{CE}}$ or $\hat{\psi}_{\text{EIS}}$, grows. For the true value, we take the univariate means and variances of $\mathbf{G}_{\text{CE}}$ and $\mathbf{G}_{\text{EIS}}$ respectively. Additionally, we perform a bias-variance decomposition to see where the estimation error originates.

More concretely, fix N and denote by $\mu, \sigma^2 \in \mathbf{R}^{(n+1)\cdot m}$ the marginal means and variances of $\mathbf{G}_{\text{CE}}$ ($\mathbf{G}_{\text{EIS}}$). Let $\hat{\mu}_i, \hat{\sigma}_i^2 \in \mathbf{R}^{(n+1)\cdot m}$ be the marginal means and variances of $\mathbf{G}^{N,i}_{\text{CE}}$ ($\mathbf{G}^{N,i}_{\text{EIS}}$) for $i = 1, \dots, M$. Now

$$\widehat{\text{aMSE}} = \frac{1}{M} \frac{1}{(n+1)m} \sum_{i=1}^{M} \|\mu - \hat{\mu}_i\|_2^2 + \|\sigma^2 - \hat{\sigma}_i^2\|_2^2$$

is an estimate of the mean-squared error of (μ, σ^2), where we divide by $(n+1)m$ to make estimates comparable across models of different dimensions.

In Figure 3.6 we show the $\widehat{\text{aMSE}}$ for both the CE-method and EIS for varying values of N. As is evident from this Figure, the CE-method consistently has a larger aMSE than EIS, for all values of N. Thus the CE-method requires several orders of magnitude more samples to obtain the same precision as EIS.

For further investigation, we perform a bias-variance decomposition of the aMSE for both the means μ and variances σ^2. Consider the average means and variances over the M simulations,

$$\bar{\mu} = \frac{1}{M} \sum_{i=1}^{M} \hat{\mu}_i \qquad\qquad \bar{\sigma}^2 = \frac{1}{M} \sum_{i=1}^{M} \hat{\sigma}_i^2,$$

and the state-average squared bias and variance

$$\begin{aligned}
\text{aBias}^2_{\mu} &= \frac{1}{(n+1)m} \|\mu - \bar{\mu}\|_2^2, \\
\text{aVar}_{\mu} &= \frac{1}{M-1} \frac{1}{(n+1)m} \sum_{i=1}^{M} \|\bar{\mu} - \mu_i\|_2^2, \\
\text{aBias}^2_{\sigma^2} &= \frac{1}{(n+1)m} \|\sigma^2 - \bar{\sigma}^2\|_2^2, \\
\text{aVar}_{\sigma^2} &= \frac{1}{M-1} \frac{1}{(n+1)m} \sum_{i=1}^{M} \|\bar{\sigma}^2 - \sigma_i^2\|_2^2.
\end{aligned}$$

These values are depicted in Figure 3.6.

interpretation of Figure 3.6, equal contribution of bias and var, not much to gain from bias correction

is bias of CEM really of this order? would expect bias usually to be of order 1/n, bias squared of order 1/n squared, so negligible compared to 1/n mse?

3.8.4 Numerical convergence

3.8.5 Performance of the optimal proposal

change EF to aEF (asymptotic EF) everywhere in this section

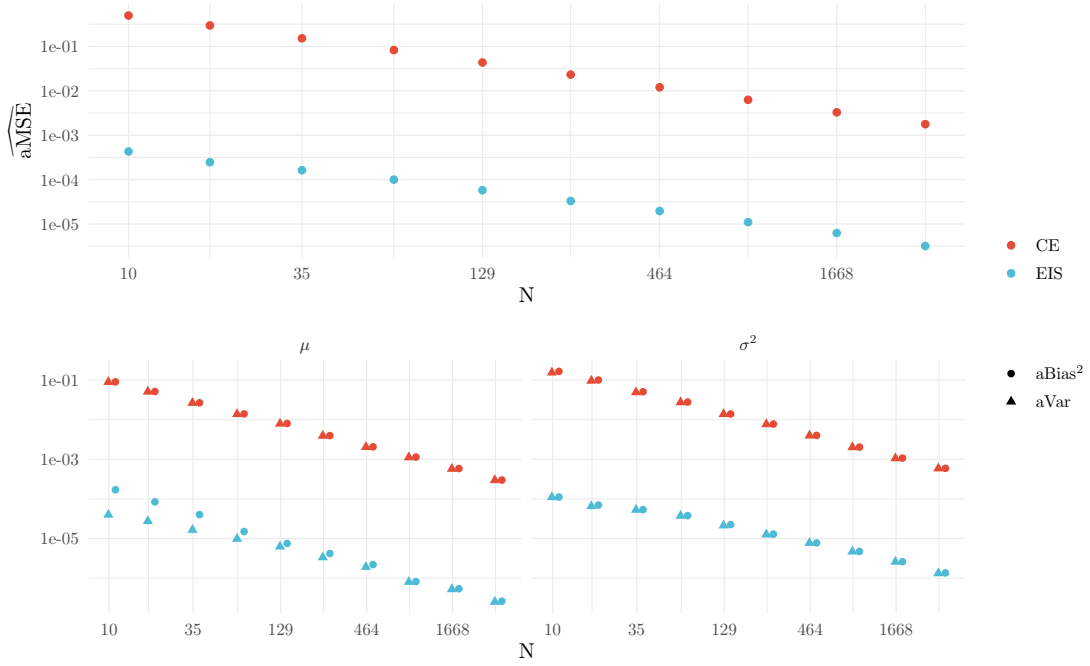

Figure 3.6: TODO

For the performance of importance sampling the efficiency factor $\mathrm{EF} = \frac{\mathrm{ESS}}{N}$ plays an important role, see Section 3.4. Additionally, it allows a comparison of the effectiveness of importance sampling across multiple sample sizes N, indeed, as $N \to \infty$, EF converges to ρ^{-1}, where ρ is the second moment of importance sampling weights, $\int w^2 \, \mathrm{d}\mathbf{G}$.

Returning to the distributions studied in Examples 3.4 and 3.5, we now calculate the asymptotic efficiency factor

$$\mathrm{EF} = \frac{1}{\rho} \in (0, 1].$$

As the proposal is always $\mathcal{N}(\mu, \sigma^2)$ with either μ or σ^2 fixed, and $\mathbf{P}$ is a mixture of Gaussians or $\mathcal{N}(0, 1)$, ρ is analytically available.

For Example 3.4, both EIS and the CE-method have, by symmetry, the same optimal $\mu = 0$. Thus the efficiency factor only depends on the fixed σ^2, see Figure 3.7, and is the same for EIS and the CE-method.

For Example 3.5 the two methods have different optimal proposals, thus also different asymptotic efficiency factors. In Figure 3.8, the first two subfigures show how the efficiency factor depends on the misspecified mu for both methods. The optimal variances are based on the results from Example 3.5, i.e. based on simulation for EIS. The right-hand subfigure shows the relative efficiency factor, i.e. the ratio of the efficiency factor for the CE-method and EIS. Here values smaller than 1 indicate that EIS has a larger efficiency factor than the CE-method.

In this figure, we can observe that, as expected, misspecification in μ almost always results in a smaller efficiency factor, an exception being the scale mixture with $\varepsilon^2 = 100$ for the CE-method. Compared to Figure 3.7, we see that already small misspecification in μ results in a large decline in EF, although we should keep in mind that this is not a fair comparison, as μ and σ^2 live on different scales. If $\mu = 0$ is correctly specified, both methods have comparable performance, except for extreme cases of the mixture models, i.e. when $\omega^2 = 0.1$ or when $\varepsilon^2 = 100$. For small misspecification of μ, this remains true, but for larger misspecification, the CE-method has a larger efficiency factor, especially for the bimodal location mixture with $\omega^2 = 0.1$, where the performance of EIS deteriorates.

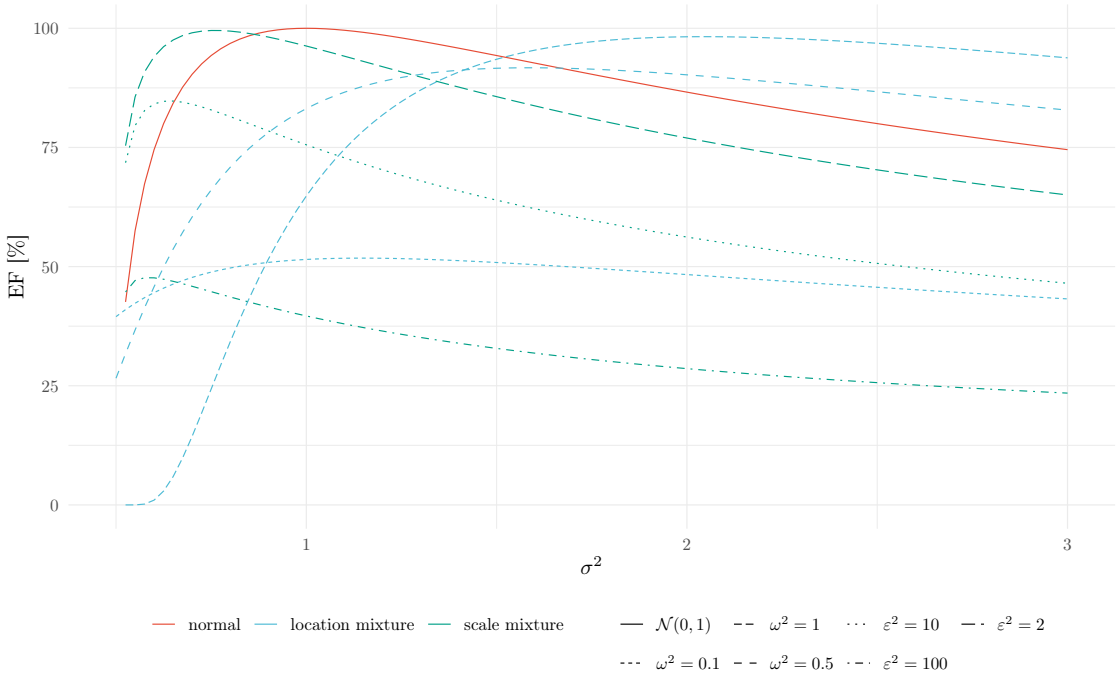

Figure 3.7: TODO

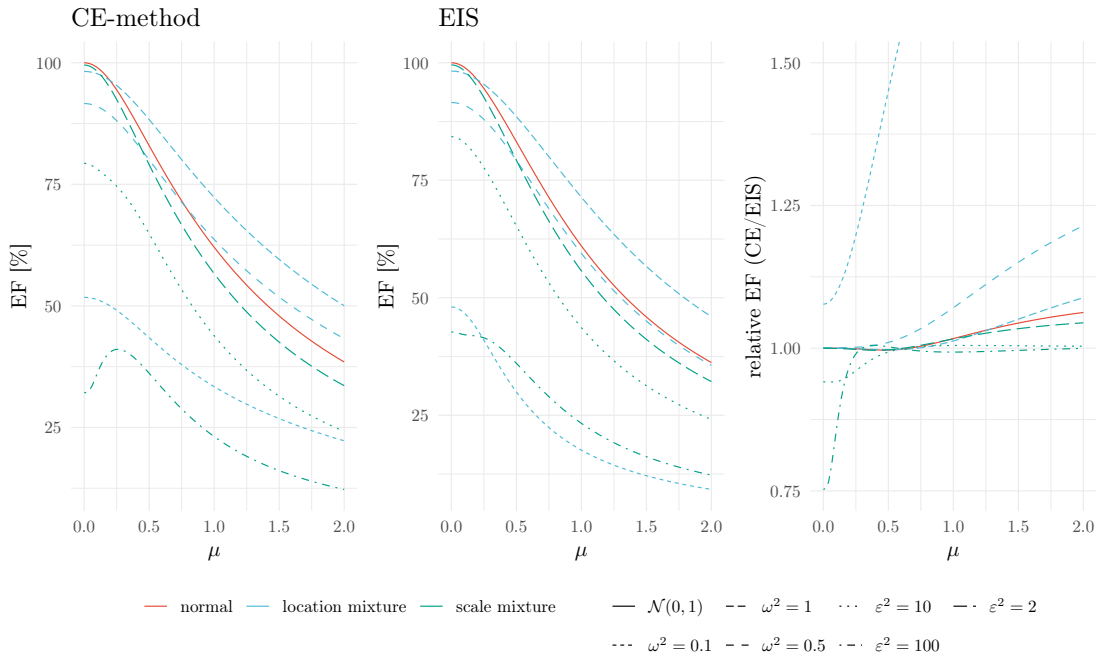

Figure 3.8: TODO

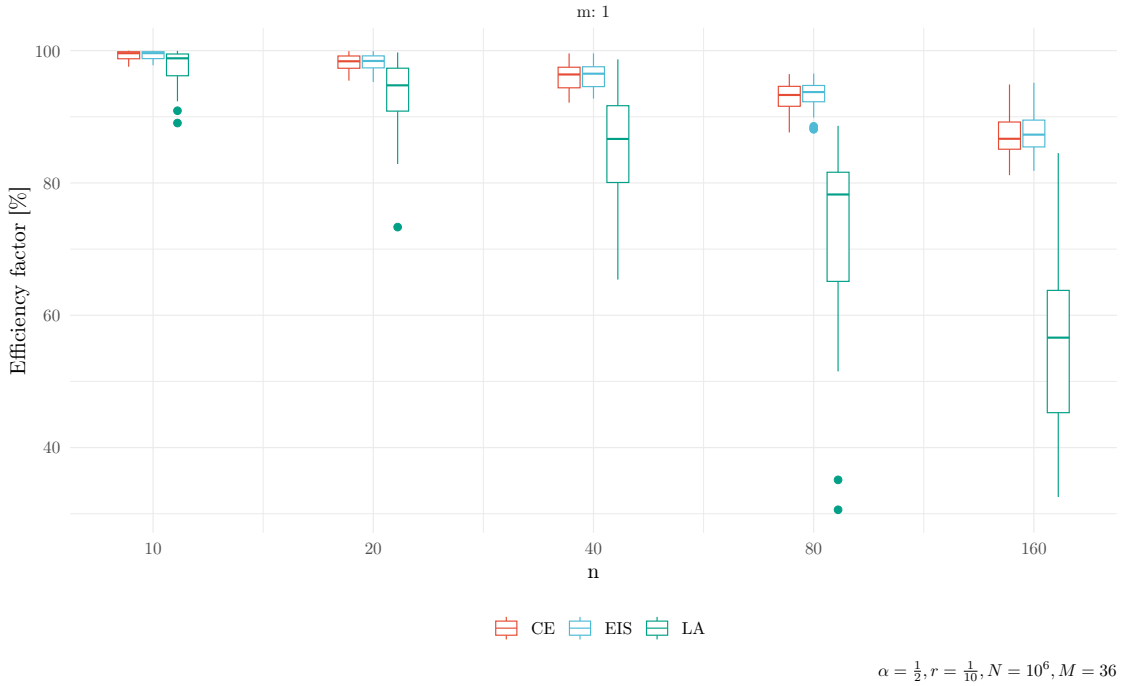

Figure 3.9: The asymptotic efficiency factor degenerates as the number of time steps n increases. We show the estimated efficiency factor over 100 replications of estimating the optimal parameters for Example 3.6 with the CE-method and EIS with $N_{\text{true}} = 10^6$ and the resulting estimated efficiency factors at the optimum. Notice the log scale of the x-axis. The performance of the optimal CE-method and EIS parameters is comparable and superior to that of the LA

stress that cem gives global optimum, eis only approximate

For the model from Example 3.6 we cannot determine ρ analytically, so we fall back to a simulation study. Thus, we also estimate EF for each of the M runs, using the same number of samples $N = N_{\text{true}}$ as was used to determine the true optimal parameter. We display the resulting efficiency factors in Figure 3.9. The parameters α, r, N, M may be found in the bottom right corner of the figure. For a low number of time steps n, all three methods perform comparably. With increasing n, their performance expectedly worsens, however, more so for the local LA, while the CE-method and EIS perform comparably around their optimal value.

3.9 Conclusion

compare independent components exponential family

Chapter 4

Analysis of selected models

Contributions of this chapter

The main contribution of this chapter is to apply the methods derived in Chapter 3 to selected inference and prediction problems in the context of COVID-19 in Germany.

Spatial reproduction number model

Regional growth factor model

4.1 Spatial reproduction number model

(i) essentially the Regional model presented in ECMI

(ii) rationale: compare to weekly GF which has "stable" regional effects

This section is based on (Burgard et al., 2021) where we fitted a similar model to daily COVID-19 incidences in Germany. However, in this previous work, we fitted a separate model for each day whereas the model presented in this section additionally regularizes in time by extending the old model to a PGSSM.

4.1.1 Context

4.1.2 Data

4.1.3 Model

Recall from Section 2.2 the stochastic renewal equation

ref it

$$I_t | I_{t-1}, \cdots \sim \text{Pois}\left(R_t \sum_{\tau=1}^{k} I_{t-\tau} w_\tau\right)$$

for the incidences at day t I_t, the time-varying reproduction number R_t and the infectivity profile $w \in \mathbf{R}^k_{\geq 0}$ with $\sum_{\tau=1}^{k} w_\tau = 1$.

4.1.4 Discussion

4.2 Regional growth factor model

4.2.1 Context

4.2.2 Data

4.2.3 Model

4.2.4 Discussion

4.3 Nowcasting hospitalizations

4.3.1 Context

Judging the severity of the COVID-19 epidemic has been an ongoing challenge since its inception. As immunization against COVID-19 rose, strict enforcement of social distancing rules eased and testing regimes became less strict, case incidences became a less reliable and harder to interpret indicator of epidemic severity. Instead more direct indicators of morbidity, such as the number of deaths and ICU admissions and occupancy have come to the fore. But these indicators are late due to the substantial delays between infection and occurence. An alternative indicator that captures the morbidity caused by COVID-19 but is earlier than the others is the number of hospitalisations of positive COVID-19 cases.

While hospitalisations occur earlier, they still come with substantial delay between the infection and subsequent admission to hospital. Additional difficulties arise due to delays in reporting, i.e. the time it takes until the hospital reports the new case to the national health authorities. The problem of accounting for delays in reporting for occurred, but not yet reported events has been termed **nowcasting**, i.e. forecasting of the indicator at time "now". Predicting the number of hospitalisations is thus a mixture of both forecasting — which reported COVID-19 cases will end up in the hospital — and nowcasting — which cases have yet to be reported — and we will use the term nowcasting in

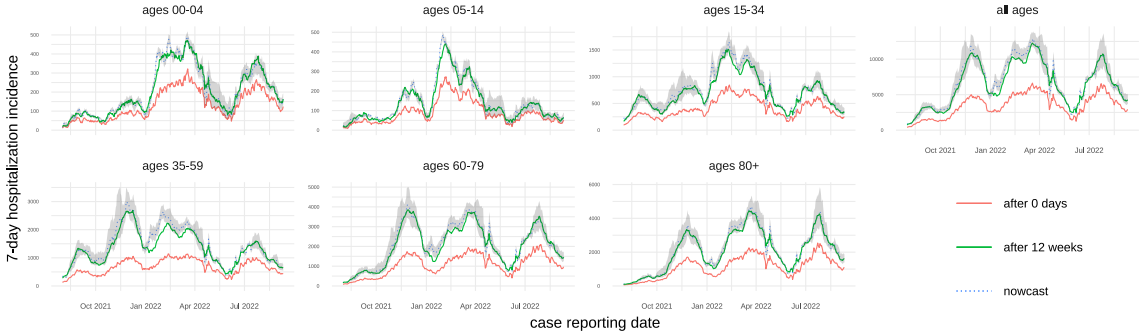

Figure 4.1: Germany's 7-day hospitalisation incidence changes due to various delays such as time to hospitalisation and delays in reporting. This figure shows the extent of these delays: incidences reported at the present date (red lines) severely underestimate the hospitalisation incidence (green solid lines) that is reported after 3 months. Our nowcasting model (blue dotted lines, 95% prediction intervals in shaded gray) deals with this problem by predicting the hospitalisation incidence based on past cases and their delays to hospitalisation.

this paper to mean this predictive mixture. In this section we focus on the situation in Germany where data on hospitalisations has been available since April 2021 provided by the German federal health care authorithy, the Robert Koch-Institut (RKI), via Github Robert Koch-Institut, 2021. In these data the number of hospitalisations is linked to the date of reporting of the associated case, so the term of nowcasting is accurate: we are interested in the "true" value of the indicator today, that will only be observed after a long delay. While this association requires a careful interpretation of the indicator (see Section 4.3.4) it was, besides case incidences and ICU occupancy, one of the main official indicators in Germany informing countermeasures in 2021 and so there is merit in nowcasting it.

The extent of delays is visible in Figure 4.1: the reported number of hospitalisations will roughly double over the course of twelve weeks. By the aforementioned reporting scheme of hospitalisations there are two reporting dates for a single hospitalised case: the reporting date of the case, i.e. the date when local health authorities were made aware of the positive test, and the reporting date of the hospitalisation, i.e. when the hospitalisation was reported to the RKI. This induces a double weekday effect in the reporting delays which we make visible in Figure 4.2.

Compared to other approaches in the COVID-19 NowcastHub, that tended to exclusively focus on modelling the delay distribution with parametric and non-parametric models, our model sidesteps this complex delay structure by decomposing delayed hospitalisations into weekly chunks (Figure 4.4) and incorporating case data. As cases and hospitalisations are explicitly linked by the case reporting date we forecast the number of hospitalisations in each chunk based on the current incidences and past fractions of hospitalisations in a comparable weekly chunk. We additionally quantify uncertainty by prediction intervals that are informed by the past performance of our model. This makes our model straightforward to understand, easy to implement and fast to run.

reformulate

The origin of nowcasting lie in accounting for incurred, but not reported claims in the actuarial sciences Kaminsky, 1987, delays in reporting for AIDS Lawless, 1994; Zeger, See, and Diggle, 1989 and other infectious diseases Farrington et al., 1996. Popular statistical approaches include methods from survival analysis Lawless, 1994 and generalized linear regression Zeger, See, and Diggle, 1989. In the survial analysis setting one commonly models the reverse time discrete hazard parametrically and assumes multinomial sampling of the final number of cases, potentially accounting for overdispersion. This has been studied with frequentist Midthune et al., 2005 and Bayesian An Der Heiden and Hamouda, 2020; Höhle and An Der Heiden, 2014 methods. The generalized linear regression approach has origins in the chain ladder model from actuarial sciences Renshaw and Verrall, 1998 and models the observed counts in the reporting triangle by a Poisson or negative binomial distribution. For both approaches, available covariates can be incorporated in a straightforward way. In the setting

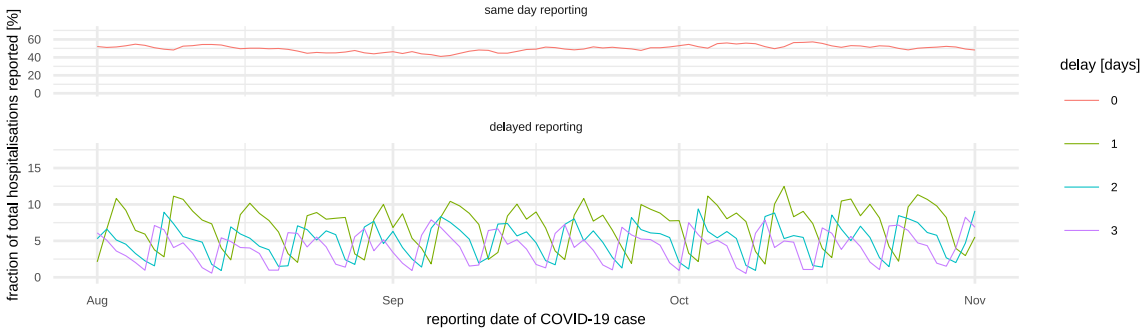

Figure 4.2: The hospitalisation incidence contains a double weekday effect owed to the reporting of both the COVID-19 case and the subsequent hospitalisation. While the weekday effect of the case reporting date is somewhat mitigated by summing over 7 day periods, the weekday effect of reporting date of the hospitalisation is still present in the data. It is most pronounced for hospitalisations that are reported with delays, i.e. where the case reporting date does not match the reporting date of the hospitalisation.

of real-time nowcasting, it is often beneficial to incorporate epidemic dynamics into the model, this can be achieved by splines Höhle and An Der Heiden, 2014; van de Kassteele et al., 2019 or by a latent process of infections McGough et al., 2020.

Nowcasting methods have wide application in accouting for reporting delays Midthune et al., 2005, early outbreak detection Bastos et al., 2019; Salmon et al., 2015, and, in the recent COVID-19 epidemic, improving real-time monitoring of epidemic outbreaks Akhmetzhanov, 2021; An Der Heiden and Hamouda, 2020; Günther et al., 2021; Schneble et al., 2021. Evaluating a forecasting model in a real-time public health setting is advantageous as it avoid hindsight bias Desai et al., 2019, however nowcasting approach may have difficulties with bias and properly calibrated uncertainty if used in a real-time setting. This includes rapidly changing dynamics Günther et al., 2021; van de Kassteele et al., 2019, both of the delay distribution and the underlying epidemic, retrospective changes in data Midthune et al., 2005 and long delays with few observed cases Noufaily et al., 2015.

To avoid the aforementioned hindsight bias one can make their predictions publicly available in real-time Bracher et al., 2021; Ray et al., 2020. For the hospitalisations in Germany, Thomas Hotz and I have participated in the German COVID-19 NowcastHub *Nowcasts Der COVID-19 Hospitalisierungsinzidenz* 2022 since November 2021 where nowcasts are available in a public Github repository *Hospitalization Nowcast Hub* 2022 with the "ILM-prop" model. The ideas, especially the model and the "double-weekday effect", discussed this section are based on this model. However, the "ILM-prop" model is based on simple point estimates for the proportion of hospitalisations per reported case, neglecting regularization over time. In this thesis we extend this model to the SSM setting of this thesis and investigate if the increased model complexity results in improved performance.

4.3.2 Data

To predict the number of hospitalisations we consider the reporting process of both reported COVID-19 cases and reported hospitalisations. Recall that the reporting date of a COVID-19 case is shared for both the case and its hospitalisation, i.e. the case and hospitalisation are linked through this date.

As hospitalisations are only available as 7-day rolling sums, we use 7-day rolling sums for daily reported incidences as well. To avoid dealing with the double weekday effect of both reporting date of the case and reporting date of the hospitalisation (see Figure 4.2) we divide the future hospitalisations we wish to predict into chunks of one week, which gets rid of the weekday effect for the hospitalisations. This is depicted in Figure 4.4. Our prediction of each of these weekly chunks then consists of the fraction of hospitalisations of reported cases in the past.

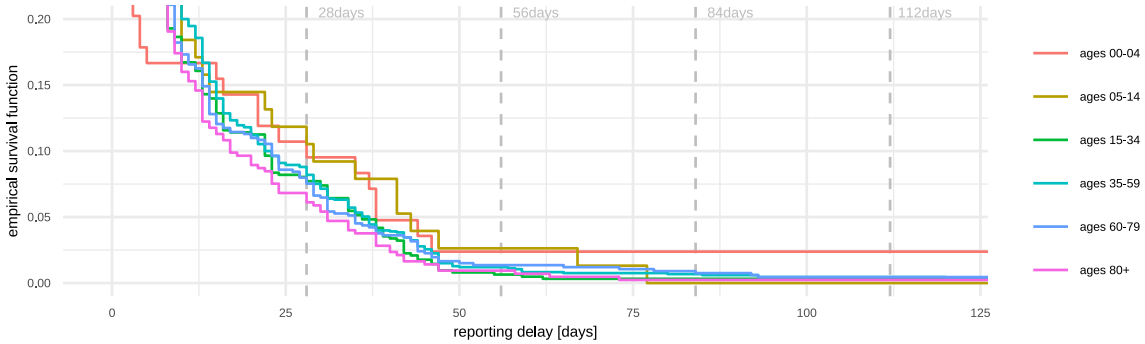

Figure 4.3: Survival function of reporting delays of weekly hospitalisations $H_{t,d}$ with case reporting date 01 September 2021. The delay distribution has long tails with a non-neglibgible fraction of observed delays longer than eight weeks, in some age groups even twelve weeks.

We use publicly available data from the German national health authority (RKI) on daily reported COVID-19 cases Robert Koch-Institut, 2022a and weekly reported hospitalisations Robert Koch-Institut, 2021. Both datasets are updated on a daily basis.

COVID-19 cases are described by their date of reporting, i.e. the date that the local health authorities were made aware of the case. For a fraction (63 %) of cases the date of symptom onset is also reported. Due to delays in the process from infection to reporting – e.g. the time it takes to get tested, evaluate the test and report the result to local health authorities – the date of reporting is, for most cases, some days after symptom onset (median delay: 3 with interquartile range $[2, 6]$). As the date of symptom onset is not known for a substantial amount of incident cases, and is not reported for hospitalised cases, we focus our analysis on the date of reporting.

Hospitalisations are associated with the *reporting date of the corresponding case* and no information is available on the actual date of hospitalisation. In addition, hospitalisations are only published as weekly sums over the past seven days. This means that the number of hospitalisations reported for today consists of all hospitalisations that correspond to cases that have a *case reporting date* in the past seven days. In particular if the case reporting date of a hospitalised case is today the case will *not* count towards todays hospitalisation count. The reporting date of hospitalisation is not available in the dataset, but can be inferred by comparing datasets from consecutive days.

Daily incident cases and weekly hospitalisations are reported by federal state and age group (00-04, 05-14, 15-34, 35-59, 60-79, 80+). Incident cases are additionally reported by county and sex.

In line with the structure of the data provided by the RKI we let $H^a_{t,d}$ be the number of weekly hospitalisations in age group a with case reporting date $t-1, \dots, t-7$ that are known on day $t+d$, aggregating over all states. Accordingly we define $I^a_{t,d}$ to be the number of weekly incident cases in age group a with reporting date $t-1, \dots, t-7$ that are known on day $t+d$. Finally we reconstruct the reporting triangles for weekly hospitalisations (Figure 4.4) by differencing the $H^a_{t,d}$ for fixed t: $h^a_{t,d} = H^a_{t,d} - H^a_{t,d-1}$, setting $H^a_{t,-1}$ to 0 by convention. We recover the reporting triangle $i^a_{t,d}$ for incident cases in the same manner.

We show the empirical surivival function of hospitalisations for a fixed date in Figure 4.3. We observe that delays have long tails, with most cases reported after 12 weeks (84 days), except for the youngest age group. After such a long delay between infection and hospitalisation we deem it unlikely that hospitalisation is due to COVID and disregard all longer delays accordingly. Given such long delays, it does not suffice to nowcast only todays hospitalisations, but also for dates in the past to monitor hospitalisation, i.e. observe current trends; we thus nowcast for all delays $d = 0, \dots, 28$.

4.3.3 Model

More formally, denote by $h_{t,d}$ the number hospitalisations with reporting date t that are known d days later. Unfortunately we only observe

$$H_{t,d} = \sum_{s=t-6}^{t} h_{s,d+(t-s)},$$

i.e. a weekly sum of reported hospitalisations. On day T our goal is to predict $H_{t,D}$ for large delays D and days $t \leq T$, of course it suffices to predict $H_{t,D} - H_{t,T-t}$ and add the known $H_{t,T-t}$ to this prediction. We rewrite this into weekly telescoping sum

$$H_{t,D} - H_{t,d} = (H_{t,d+7} - H_{t,d}) + (H_{t,d+14} - H_{t,d+7}) + \cdots + (H_{t,D} - H_{t,d+7K}),$$

where $K = \lfloor (D-d)/7 \rfloor$, reducing the task at hand to predict hospitalisations in the k-th week ahead, $H_{t,d+7k} - H_{t,d+7\cdot(k-1)}$, $k = 1, \ldots, K$. To leverage known reported incidences, rewrite this as

$$\underbrace{\frac{H_{t,d+7k} - H_{t,d+7\cdot(k-1)}}{I_{t,d}}}_{=:p_{t,d,k}} I_{t,d}$$

where $I_{t,d}$ is the 7-day case incidence with reporting date t known at time $t+d$, i.e. the incidenct case analouge of $H_{t,d}$.

Assuming that the proportions $p_{t,d,k}$ change slowly over time t we estimate them by

$$\widehat{p_{t,d,k}} = \frac{H_{t-7k,d+7k} - H_{t-7k,d+7\cdot(k-1)}}{I_{t-7k,d}} = p_{t-7k,d,k} \tag{4.1}$$

and finally predict

$$\widehat{H_{t,D}} = H_{t,d} + I_{t,d} \left(\widehat{p_{t,d,1}} + \cdots + \widehat{p_{t,d,K}} \right). \tag{4.2}$$

As hospitalisation is affected by age, we perform this procedure for all available age groups separately and finally aggregate over all age groups to obtain a nowcast for all age groups combined.

This describes our point nowcast for 7-day hospitalisations. To obtain uncertainty intervals we fit a normal (age groups 00-04 and 05-14) or lognormal (all other age groups) distribution to the past performance of our model. We chose these distributions based on explorative analysis and believe that these should be seen as heuristics rather than as a matter of fact, which is in line with the philosophy of our model to be as simple as possible.

Denote by $\hat{H}_{t,D,s}$ the nowcast made for date t on date $s \geq t$. Starting with date $t+D$ the definite $H_{t,D}$ is known and we can estimate the absolute prediction error $\varepsilon_{t,s} = H_{t,D} - \hat{H}_{t,D,s}$ and the relative prediction error $\eta_{t,s} = \log(H_{t,D} - H_{t,s-t}) - \log\left(\hat{H}_{t,D,s} - H_{t,s-t}\right)$. For the nowcast for date t made on date s we estimate the standard deviation $\hat{\sigma}$ of $\varepsilon_{t-D-i,s-D-i}$ or $\eta_{t-D-i,s-D-i}$ (age groups 00-04, 05-14 and others respectively), $i = 0, \ldots, 27$ by its empirical counterpart. The estimated predictive distribution which informs our prediction intervals is then $\mathcal{N}(\hat{H}_{t,D,s}, \sigma^2)$ (age groups 00-04 and 05-14) or $\mathcal{LN}\left(\log\left(\hat{H}_{t,D,s} - H_{t,s-t}\right), \sigma^2\right) + H_{t,s-t}$ (all other agr groups).

4.3.4 Discussion

Before evaluating the predictive performance of our model we investigate how the fraction of hospitalisations after one up to four weeks changes over time across different age groups. Figure 4.5 shows that these fractions are changing slowely over time, especially in the older age groups. Due

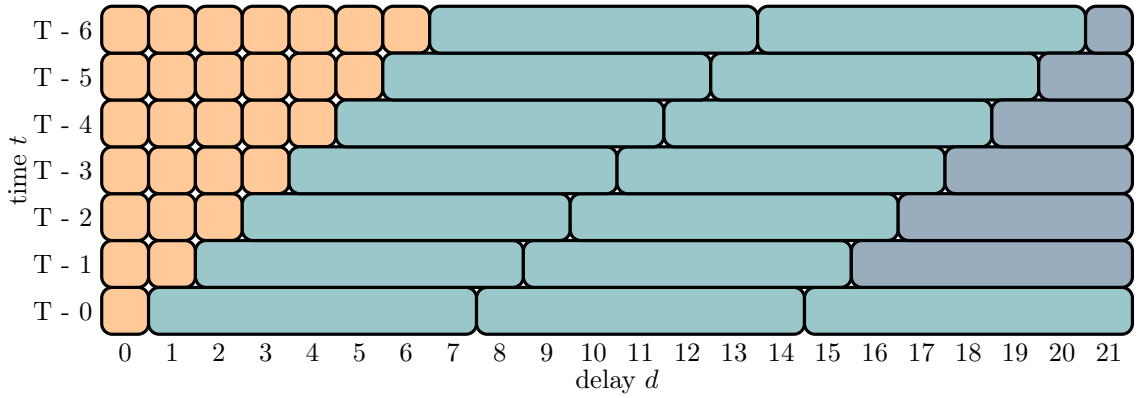

Figure 4.4: Decomposition of the daily reported hospitalisation incidences into the known incidences, i.e. the **reporting triangle**, and the future weekly increments. The last increment might not be a weekly one, but we expect few cases to occur for such long delays.

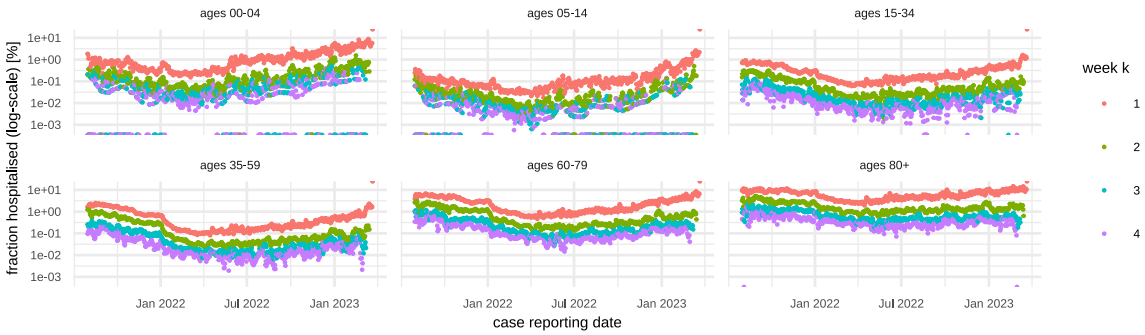

Figure 4.5: We show the fractions of hospitalisations in the k-th week after case reporting date t of initially reported cases in different age groups, i.e. $p_{t,0,k} = \left(H_{t,7k} - H_{t,7\cdot(k-1)}\right)/I_{t,0}$. Note the log-scale of the y-axis. During periods of low incidence, e.g. July – September, we find large fluctations, but no discernable weekly pattern. With rising case numbers the fractions stabilise and decrease in most age groups. This might be due to changes in testing regime detecting less severe cases. As changes occur on slow time scales, estimating these fractions by Eq. (4.1) is a promising approach.

to smaller numbers of infections and hospitalisations reported in the younger age groups these fractions vary more strongly, occassionally dropping to 0. Across all age groups we observe a steady decline from October 2021 to December 2021 with a steeper drop in fraction of hospitalisations starting with January 2022. The former period corresponds to a time of mandatory testing at the workplace which may improve ascertainment of asymptomatic and less severe cases. The latter effect is most recognizable in the 35-59 age group and coincides with the time that the Omicron variant became dominant in Germany Robert Koch-Institut, 2022b. Additionally there is no visually discernible weekday effect present in Figure 4.5.

In Figure 4.1 we depict the nowcasts produced from our model including 95% prediction intervals, whose lengths are based on the past performance of our model. Except for the period from January to April 2022, the model produces reasonable nowcasts with prediction intervals that have sensible widths. In the aforementioned period the nowcasts overpredict the final hospitalisations, except for the oldest age group, and, after a transitionary period, have larger uncertainty.

To investigate the quality of point predictions we display the time-evolution of absolute (AEP) and relative errors of predictions (REP, $\log_{10}$-scale) across all age groups in Figure 4.6. From this figure one can infer that the point nowcasts produced by our model tend to slightly overpredict the final number of hospitalisations. Indeed, the interquartile range of REPs for all age groups and dates combined spans $[-1.56, 8.33]$, demonstrating the same tendency. The highest REPs occured

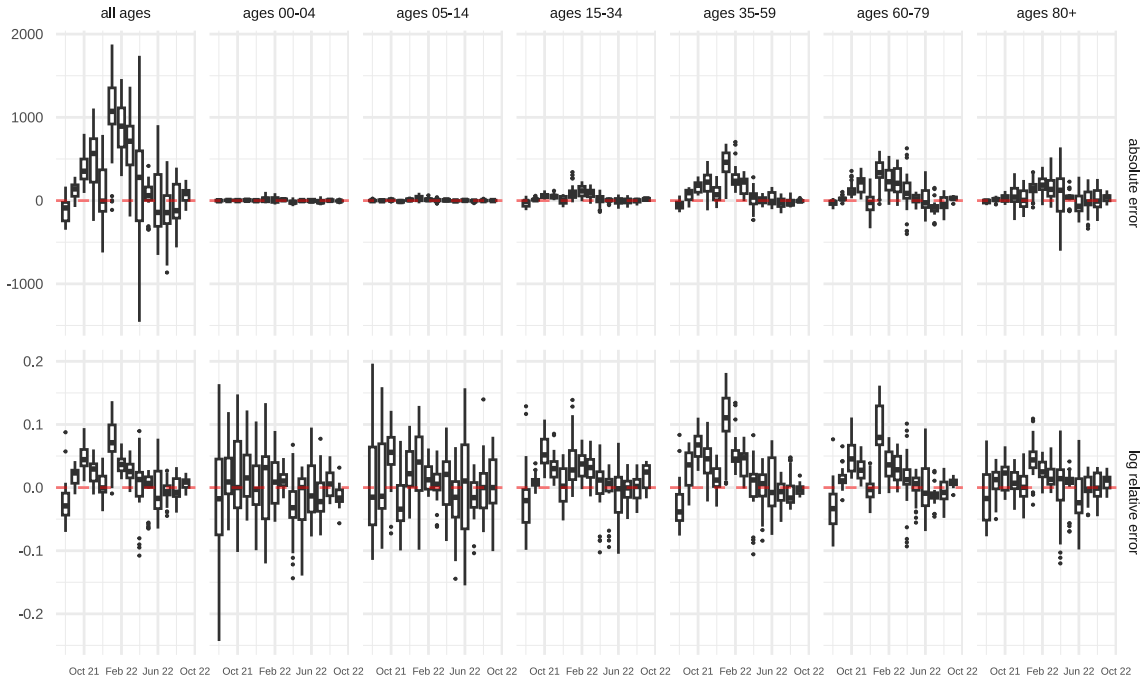

Figure 4.6: We show relative and absolute errors of prediction of our model for same day nowcasts by month of forecast and selected age groups. Relative errors are displayed on the log10 scale, i.e. as $\log_{10}(\text{predicted}) - \log_{10}(\text{actual})$. Up to December 2021 the model performs well, especially in the older age groups where most hospitalisations occur. The sharp increase in cases in January 2022 coupled with a lower probability of hospitalisation, most likely due to the appearance of the Omicron variant in Germany, lead to overpredictions across all age groups.

in October / November 2021 and January/February 2022; the first corresponding to introduction of mandatory testing at the workplace and the second to the arrival of the Omicron variant in Germany. In both circumstances the number of cases rose while hospitalisations did not increase proportionally, a similar effect to the one observed in Figure 4.5.

We further quantified uncertainties in estimaton by uncertainty intervals based on an assumption of a (log)-normal distribution for the errors with standard deviation based on past performance of our model. In Figure 4.7 we show the coverages of the 50% and 95% prediction interval across all age groups and delays for the whole time period of our study. For most age groups the 50% prediction interval has close to nominal coverage, while the 95% intervals have less than nominal coverage.

As our goal is to capture all of the uncertainty in this prediction, we chose to assume a sensible distribution for the prediction, a normal distribution for the two young age groups and a log-normal distribution for all other age groups. This has the advantage of producing more honest, wider, prediction intervals than those based on parametric distributions. The estimated standard deviation will also account for periods of low coverage, such as January 2022, albeit only after the maximum delay of 12 weeks.

We base our choice of 12 weeks of delay on the empirical survival function displayed in Figure 4.3. One could, however, argue for shorter maximum delay such as 6 weeks because time from reported infection to hospitalisation is much shorter, on the order of $\approx$ 10 days Faes et al., 2020, so hospitalisations after this (shorter) period are unlikely to be due to the acute infection with SARS-CoV-2. This would have two main benefits: The model would adapt faster to changing circumstances and the indicator nowcasted describes the severity of the epidemic more appropriately.

The main advantage of our model over established nowcasting approches is its simplicity, making it easy to understand, straightforward to implement and, once the reporting triangles for incidences and hospitalisation are created, fast to run; taking only $< XX$ minutes on a standard notebook(**TODO:**

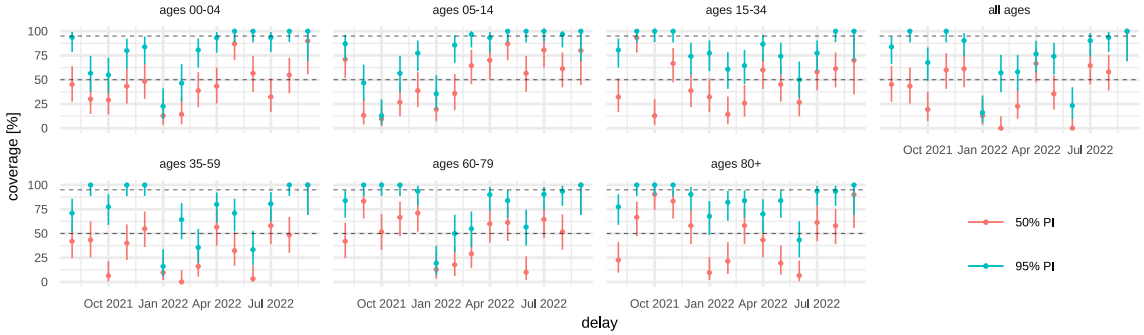

Figure 4.7: Empirical coverage of 50% and 95% prediction intervals (PI) based on same-day nowcasts for dates 2021-08-01 to 2022-09-10 (406 dates) for which the true amount of hospitalisations after 12 weeks is known as of the writing of this paper. We also display pointwise 95% binomial confidence intervals for the coverage. Given the difficulties of real-time forecasting Desai et al., 2019 we deem the coverages good, except for the transitionary period in the end of 2021 where changing testing schemes and the change from Delta to Omicron cause our model to be overly confident. Coverage is generally better in the older age groups.

check!).

The problem of nowcasting hospitalisations is different from previously studied nowcasting settings in several ways. At the time of nowcast a large fraction of hospitalisations are not only unobserved, but are yet to occur - in this sense the nowcast is more accurately termed a forecast. As the date of hospitalisation is not known, the hospitalisations are associated by the date of reporting of the COVID-19 case, creating the double-weekday effect displayed in Figure 4.2. While daily updated data on hospitalisations are available, these consist only of moving weekly aggregates, consecutive observations are strongly auto-correlated.

We sidestep all of these issues by splitting the hospitalisations to nowcast into weekly chunks, incorporating leading indicators of hospitalisation – the weekly reported case incidences – and modelling the number of hospitalisations to come in each chunk by binomial thinning of incidences. Let us stress that this approach is only possible in the special situation where case and hospitalisation are explicitly linked, however we believe that incorporating leading indicators into nowcasting models is a promising approach.

An additional advantage of our model is that the hospitalisation probabilities can further be analysed, e.g. by investigating association between the publicly available vaccination rates and the probability of hosptialisation and delay to hospitalisation. Sudden changes in these fractions, as observed in Figure 4.5, can also hint towards worse model performance, especially if this change can be attributed to changing probability of hospitalisation due to new variants or changing testing regime.

Real time forecasting of epidemiological indicators is a difficult task Desai et al., 2019, in particular quantifying uncertainty Bracher et al., 2021. To test our model under real-time circumstances we submitted daily nowcasts to the German COVID-19 NowcastHub *Nowcasts Der COVID-19 Hospitalisierungsinzidenz* 2022 since Novemer 2021. In the nowcasting context, Lawless, 1994 goes to great lengths to account for overdispersion due to changes in delay distribution, introducing gamma and Dirichlet priors and explicitly modelling trends. Such an approach would also be feasible for our approach, e.g. model incidences by an appropriate Poisson or negative binomial distribution and, conditional on incidences, model hospitalisations by a binomial distribution. As this increases the complexity of our model and relies on the assumed distributions being sensible we opted for another approach.

Regarding the indicator we stress that its value on a given date does not represent the current occupancy of hospitals in Germany with COVID-19 patients but is rather an approximation to the morbidity caused by COVID-19 on that date. The reason for this discrepancy is that hospitalisations are attributed to the reporting date of the associated case, not that of hospitalisation. While the

reporting date of the hospitalisation can be recovered from the publicly available data, the date of hospitalisation cannot. Additionally, no information on the duration of stay is available, making it impossible to create an indicator for the occupancy of hospitals based solely on data provided by the RKI.

Implicit in all of these approaches is an assumption of "stationarity", i.e. that future reported hospitalisations will behave as they did in the past. Thus, all of these approaches might still be insufficient if circumstances change drastically, for example introduction of new testing schemes (school, 3G at workplace), changes in the delay distribution due to new variants, or hospitals close to capacity taking longer to process cases.

In summary, because models usually only capture a small part of the highly dynamic data-generating process, we believe that uncertainty in such circumstances should not come from unrealistic parametric assumptions but rather be based on past model performance. Given the discussed difficulties and the changing epidemiological dynamics in the period studied, the observed errors of prediction (Figure 4.6) and coverages of prediction intervals (Figure 4.7) are satisfying.

In this paper we provide a straight-forward model for nowcasting hospitalisations associated with COVID-19 in Germany. By leveraging known incident cases, we can estimate fractions of hospitalisations in weekly chunks which in turn avoids a complicated model of the two weekday effects present in the data. As the circumstances of the epidemic are changing constantly, e.g. vaccination coverage, testing regimes and emerging variants, we based uncertainty not on parametric assumptions but on the past performance of our model, assuming a (log)normal predictive distribution. We contributed nowcasts based on this model since November 2021 to the German COVID-19 NowcastHub *Nowcasts Der COVID-19 Hospitalisierungsinzidenz* 2022, a collaborative platform collecting and aggregating such nowcasts from multiple research groups. The performance of the nowcasts in this Hub and presented in this paper (Figure 4.6 and Figure 4.7) are, regarding the simplicity of the model and the highly dynamic situation, quite satisfying.

There are multiple extensions to our model worth investigating. Firstly hospitalisations are also available at the federal state level so nowcasting on a spatial scale is naturally of interest due to hetereogeneity in immunisation status and testing regimes across states. However, splitting hospitalisations into six age groups and 16 federal states will result in small numbers with larger variability which in turn increases variabliity in estimates $p_{t,d,k}$ and thus predictions which may require some regularisation thus increasing the models complexity. Secondly, in a similar vein, modeling the temporal evolution of hospitalisation probabilities by smooth functions, e.g. splines Schneble et al., 2021; van de Kassteele et al., 2019, may help in early detection of changing circumstances and thus lead to better forecasts. Thirdly our uncertainty intervals account for the variance of past performance, but Figure 4.6 suggests that there is substantial bias in periods of changing circumstances which could be incorporated into our model in a straightforward way. Finally to predict the future course of the epidemic forecasting hospitalisations for dates t that lie in the future is of interest, which could be accomplished if one has a model that produces forecasts for incidences for all age groups.

Chapter 5

Discussion

Appendix A

Implementation in Python

Appendix B

Additional calculations

Equation (3.10)

To show Equation (3.10) we calculate the second moment of $w(X)X$,

$$\begin{aligned}
\mathbb{E}(w(X)X)^2 &= \int w(x)^2 x^2 g(x)\,\mathrm{d}x \\
&= \int \sigma^2 \exp\left(-x^2\left(1-\frac{1}{\sigma^2}\right)\right) x^2 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right)\,\mathrm{d}x \\
&= \int \sigma x^2 \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\left(2-\frac{2}{\sigma^2}+\frac{1}{\sigma^2}\right)\right)\,\mathrm{d}x \\
&= \int \sigma x^2 \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\frac{2\sigma^2-1}{\sigma^2}\right)\,\mathrm{d}x \\
&= \tau\sigma \int x^2 \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{x^2}{2\tau^2}\right)\,\mathrm{d}x \\
&= \tau^3\sigma = \frac{\sigma^4}{(2\sigma^2-1)^{\frac{3}{2}}}
\end{aligned}$$

where $\tau^2 = \frac{\sigma^2}{2\sigma^2-1}$.

Bibliography

Abbott, S. et al. (June 1, 2020). "Estimating the Time-Varying Reproduction Number of SARS-CoV-2 Using National and Subnational Case Counts." In: *Wellcome Open Res* 5, p. 112. ISSN: 2398-502X. DOI: 10.12688/wellcomeopenres.16006.1.

Adamik, B. et al. (May 5, 2020). *Mitigation and Herd Immunity Strategy for COVID-19 Is Likely to Fail*. DOI: 10.1101/2020.03.25.20043109. Pre-published.

Agapiou, S. et al. (Jan. 14, 2017). *Importance Sampling: Intrinsic Dimension and Computational Cost*. DOI: 10.48550/arXiv.1511.06196. arXiv: 1511.06196 [stat]. Pre-published.

Akhmetzhanov, A. R. (2021). "Estimation of Delay-Adjusted All-Cause Excess Mortality in the USA: March-December 2020." In: *Epidemiology and Infection*. DOI: 10.1017/s0950268821001527. pmid: 34210370.

An Der Heiden, M. et al. (Apr. 22, 2020). "Schätzung Der Aktuellen Entwicklung Der SARS-CoV-2-Epidemie in Deutschland – Nowcasting." In: *Epidemiologisches Bulletin*. DOI: 10.25646/6692.4.

Arratia, R. et al. (1990). "Poisson Approximation and the Chen-Stein Method." In: *Statistical Science* 5.4, pp. 403–424. ISSN: 0883-4237. JSTOR: 2245366. URL: https://www.jstor.org/stable/2245366 (visited on 01/11/2024).

Arroyo-Marioli, F. et al. (Jan. 13, 2021). "Tracking R of COVID-19: A New Real-Time Estimation Using the Kalman Filter." In: *PLOS ONE* 16.1. DOI: 10.1371/journal.pone.0244474. pmid: 33439880.

Assimakis, N. et al. (2012). "Information Filter and Kalman Filter Comparison: Selection of the Faster Filter." In: *Information Engineering*. Vol. 2. 1, pp. 1–5. URL: http://madam.users.uth.gr/papers/3%20IJIE_2012.pdf (visited on 06/24/2024).

Bastos, L. S. et al. (2019). "A Modelling Approach for Correcting Reporting Delays in Disease Surveillance Data." In: *Statistics in Medicine*. DOI: 10.1002/sim.8303.

Bazaraa, M. S. et al. (2006). *Nonlinear Programming: Theory and Algorithms*. 3. ed. Hoboken, NJ: Wiley-Interscience. 853 pp. ISBN: 978-0-471-48600-8.

Bengtsson, T. et al. (2008). "Curse-of-Dimensionality Revisited: Collapse of the Particle Filter in Very Large Scale Systems." In: *Probability and statistics: Essays in honor of David A. Freedman* 2, pp. 316–334. DOI: 10.1214/193940307000000518.

Billingsley, P. (1995). *Probability and Measure*. 3rd ed. Wiley Series in Probability and Mathematical Statistics. New York: Wiley. 593 pp. ISBN: 978-0-471-00710-4.

Biswas, M. et al. (Dec. 9, 2020). "Association of Sex, Age, and Comorbidities with Mortality in COVID-19 Patients: A Systematic Review and Meta-Analysis." In: *Intervirology* 64.1, pp. 36–47. ISSN: 0300-5526. DOI: 10.1159/000512592.

Borchering, R. K. et al. (Sept. 1, 2023). "Public Health Impact of the U.S. Scenario Modeling Hub." In: *Epidemics* 44, p. 100705. ISSN: 1755-4365. DOI: 10.1016/j.epidem.2023.100705.

Bracher, J. et al. (Aug. 27, 2021). "A Pre-Registered Short-Term Forecasting Study of COVID-19 in Germany and Poland during the Second Wave." In: *Nature Communications* 12.1 (1), p. 5173. ISSN: 2041-1723. DOI: 10.1038/s41467-021-25207-0.

Bracher, J. et al. (Oct. 31, 2022). "National and Subnational Short-Term Forecasting of COVID-19 in Germany and Poland during Early 2021." In: *Communications Medicine* 2.1 (1), pp. 1–17. ISSN: 2730-664X. DOI: 10.1038/s43856-022-00191-8.

Branch, M. A. et al. (Jan. 1999). "A Subspace, Interior, and Conjugate Gradient Method for Large-Scale Bound-Constrained Minimization Problems." In: *SIAM J. Sci. Comput.* 21.1, pp. 1–23. ISSN: 1064-8275, 1095-7197. DOI: 10.1137/S1064827595289108.

Brauner, J. M. et al. (Feb. 19, 2021). "Inferring the Effectiveness of Government Interventions against COVID-19." In: *Science* 371.6531, eabd9338. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.abd9338.

Britton, T. et al. (2019). *Stochastic Epidemic Models with Inference*. Ed. by T. Britton et al. Springer.

Brooks, S. et al., eds. (2011). *Handbook for Markov Chain Monte Carlo*. Boca Raton: Taylor & Francis. 592 pp. ISBN: 978-1-4200-7941-8.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory*. Lecture Notes-Monograph Series v. 9. Hayward, Calif: Institute of Mathematical Statistics. 283 pp. ISBN: 978-0-940600-10-2.

Burgard, J. P. et al. (Aug. 31, 2021). "Regional Estimates of Reproduction Numbers with Application to COVID-19." arXiv: 2108.13842 [stat]. URL: http://arxiv.org/abs/2108.13842 (visited on 09/30/2021).

Byambasuren, O. et al. (Dec. 31, 2020). "Estimating the Extent of Asymptomatic COVID-19 and Its Potential for Community Transmission: Systematic Review and Meta-Analysis." In: *Journal of the Association of Medical Microbiology and Infectious Disease Canada* 5.4, pp. 223–234. DOI: 10.3138/jammi-2020-0030.

Chan, J. C. C. et al. (Sept. 1, 2012). "Improved Cross-Entropy Method for Estimation." In: *Stat Comput* 22.5, pp. 1031–1040. ISSN: 1573-1375. DOI: 10.1007/s11222-011-9275-7.

Chan, J. C. C. et al. (May 1, 2012). *Marginal Likelihood Estimation with the Cross-Entropy Method*. DOI: 10.2139/ssrn.2055042. Pre-published.

Chan, S. et al. (Feb. 1, 2021). "Count Regression Models for COVID-19." In: *Physica A: Statistical Mechanics and its Applications* 563, p. 125460. ISSN: 0378-4371. DOI: 10.1016/j.physa.2020.125460.

Chatterjee, S. et al. (Apr. 1, 2018). "The Sample Size Required in Importance Sampling." In: *Ann. Appl. Probab.* 28.2. ISSN: 1050-5164. DOI: 10.1214/17-AAP1326.

Chopin, N. et al. (2020). *An Introduction to Sequential Monte Carlo*. Springer Series in Statistics. Cham, Switzerland: Springer. 378 pp. ISBN: 978-3-030-47844-5.

Cori, A. (2021). *EpiEstim: Estimate Time Varying Reproduction Numbers from Epidemic Curves*. manual. URL: https://CRAN.R-project.org/package=EpiEstim.

Cover, T. M. et al. (2006). *Elements of Information Theory*. 2nd ed. Hoboken, N.J: Wiley-Interscience. 748 pp. ISBN: 978-0-471-24195-9.

Desai, A. N. et al. (Aug. 2019). "Real-Time Epidemic Forecasting: Challenges and Opportunities." In: *Health Security* 17.4, pp. 268–275. ISSN: 2326-5094. DOI: 10.1089/hs.2019.0022.

Diekmann, O. et al. (2013). *Mathematical Tools for Understanding Infectious Diseases Dynamics*. Princeton Series in Theoretical and Computational Biology. Princeton: Princeton University Press. 502 pp. ISBN: 978-0-691-15539-5.

Du, Z. et al. (July 1, 2022). "Reproduction Numbers of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Variants: A Systematic Review and Meta-analysis." In: *Clinical Infectious Diseases* 75.1, e293–e295. ISSN: 1058-4838. DOI: 10.1093/cid/ciac137.

Durbin, J. et al. (2012). *Time Series Analysis by State Space Methods*. 2nd ed. Oxford Statistical Science Series 38. Oxford: Oxford University Press. 346 pp. ISBN: 978-0-19-964117-8.

Durbin, J. et al. (Sept. 1, 1997). "Monte Carlo Maximum Likelihood Estimation for Non-Gaussian State Space Models." In: *Biometrika* 84.3, pp. 669–684. ISSN: 0006-3444. DOI: 10.1093/biomet/84.3.669.

– (2002). "A Simple and Efficient Simulation Smoother for State Space Time Series Analysis." In: *Biometrika* 89.3, pp. 603–616. DOI: 10.1093/biomet/89.3.603.

Ehre, M. et al. (Mar. 31, 2023). "Certified Dimension Reduction for Bayesian Updating with the Cross-Entropy Method." In: *SIAM/ASA J. Uncertainty Quantification* 11.1, pp. 358–388. DOI: 10.1137/22M1484031.

Engbert, R. et al. (Dec. 8, 2020). "Sequential Data Assimilation of the Stochastic SEIR Epidemic Model for Regional COVID-19 Dynamics." In: *Bull Math Biol* 83.1, p. 1. ISSN: 1522-9602. DOI: 10.1007/s11538-020-00834-8.

Engel, M. et al. (Jan. 15, 2023). "Bayesian Updating and Marginal Likelihood Estimation by Cross Entropy Based Importance Sampling." In: *Journal of Computational Physics* 473, p. 111746. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2022.111746.

Evensen, G. (1994). "Sequential Data Assimilation with a Nonlinear Quasi-Geostrophic Model Using Monte Carlo Methods to Forecast Error Statistics." In: *Journal of Geophysical Research: Oceans* 99.C5, pp. 10143–10162. ISSN: 2156-2202. DOI: 10.1029/94JC00572.

Faes, C. et al. (Jan. 2020). "Time between Symptom Onset, Hospitalisation and Recovery or Death: Statistical Analysis of Belgian COVID-19 Patients." In: *International Journal of Environmental Research and Public Health* 17.20 (20), p. 7560. ISSN: 1660-4601. DOI: 10.3390/ijerph17207560.

Farrington, C. P. et al. (1996). "A Statistical Algorithm for the Early Detection of Outbreaks of Infectious Disease." In: *Journal of The Royal Statistical Society Series A-statistics in Society*. DOI: 10.2307/2983331.

Flaxman, S. et al. (Aug. 2020). "Estimating the Effects of Non-Pharmaceutical Interventions on COVID-19 in Europe." In: *Nature* 584.7820, pp. 257–261. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2405-7. pmid: 32512579.

Fraser, C. (Aug. 22, 2007). "Estimating Individual and Household Reproduction Numbers in an Emerging Epidemic." In: *PLoS ONE* 2.8. Ed. by A. Galvani, e758. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0000758.

Fraser, D. et al. (Aug. 1969). "The Optimum Linear Smoother as a Combination of Two Optimum Linear Filters." In: *IEEE Trans. Automat. Contr.* 14.4, pp. 387–390. ISSN: 0018-9286. DOI: 10.1109/TAC.1969.1099196.

Frühwirth-Schnatter, S. (1994). "Data Augmentation and Dynamic Linear Models." In: *Journal of Time Series Analysis* 15.2, pp. 183–202. ISSN: 1467-9892. DOI: 10.1111/j.1467-9892.1994.tb00184.x.

"Discrete Spatial Variation" (2010). In: *Handbook of Spatial Statistics*. Ed. by A. E. Gelfand et al. CRC Press. ISBN: 978-0-429-13650-4.

Günther, F. et al. (2021). "Nowcasting the COVID-19 Pandemic in Bavaria." In: *Biometrical Journal* 63.3, pp. 490–502. ISSN: 1521-4036. DOI: 10.1002/bimj.202000112.

Gupta, S. et al. (Nov. 2020). *Mandated and Voluntary Social Distancing During The COVID-19 Epidemic: A Review*. DOI: 10.3386/w28139. National Bureau of Economic Research: 28139. Pre-published.

Haberman, S. J. (1989). "Concavity and Estimation." In: *The Annals of Statistics* 17.4, pp. 1631–1661. ISSN: 0090-5364. DOI: 10.1214/aos/1176347385. JSTOR: 2241655.

Heyder, S. et al. (2023). "Measures of COVID-19 Spread." In: *Covid-19 pandisziplinär und international: Gesundheitswissenschaftliche, gesellschaftspolitische und philosophische Hintergründe*. Ed. by A. Kraemer et al. Medizin, Kultur, Gesellschaft. Wiesbaden: Springer Fachmedien, pp. 51–66. ISBN: 978-3-658-40525-0. DOI: 10.1007/978-3-658-40525-0_3.

Höhle, M. et al. (2014). "Bayesian Nowcasting during the STEC O104:H4 Outbreak in Germany, 2011." In: *Biometrics* 70.4, pp. 993–1002. ISSN: 1541-0420. DOI: 10.1111/biom.12194.

Homem-de-Mello, T. (July 20, 2007). "A Study on the Cross-Entropy Method for Rare-Event Probability Estimation." In: *INFORMS Journal on Computing*. DOI: 10.1287/ijoc.1060.0176.

Hospitalization Nowcast Hub (Oct. 31, 2022). KITmetricslab. URL: https://github.com/KITmetricslab/hospitalization-nowcast-hub (visited on 11/09/2022).

Hotz, T. et al. (Apr. 18, 2020). "Monitoring the Spread of COVID-19 by Estimating Reproduction Numbers over Time." arXiv: 2004.08557 [q-bio, stat]. URL: http://arxiv.org/abs/2004.08557 (visited on 07/20/2020).

Hughes, T. D. et al. (Mar. 18, 2023). "The Effect of SARS-CoV-2 Variant on Respiratory Features and Mortality." In: *Sci Rep* 13.1, p. 4503. ISSN: 2045-2322. DOI: 10.1038/s41598-023-31761-y.

Jazwinski, A. H. (1970). *Stochastic Processes and Filtering Theory*. New York: Academic Press. 376 pp.

Johnson, N. L. et al. (1994). *Continuous Univariate Distributions*. 2nd ed. Wiley Series in Probability and Mathematical Statistics. New York: Wiley. 2 pp. ISBN: 978-0-471-58495-7 978-0-471-58494-0.

Julier, S. J. et al. (July 28, 1997). "New Extension of the Kalman Filter to Nonlinear Systems." In: Signal Processing, Sensor Fusion, and Target Recognition VI. Vol. 3068. International Society for Optics and Photonics, pp. 182–194. DOI: 10.1117/12.280797.

Jungbacker, B. et al. (Dec. 1, 2007). "Monte Carlo Estimation for Nonlinear Non-Gaussian State Space Models." In: *Biometrika* 94.4, pp. 827–839. ISSN: 0006-3444. DOI: 10.1093/biomet/asm074.

Kaminsky, K. S. (Apr. 1, 1987). "Prediction of IBNR Claim Counts by Modelling the Distribution of Report Lags." In: *Insurance Mathematics & Economics* 6.2, pp. 151–159. DOI: 10.1016/0167-6687(87)90024-2.

Kappen, H. J. et al. (Mar. 1, 2016). "Adaptive Importance Sampling for Control and Inference." In: *J Stat Phys* 162.5, pp. 1244–1266. ISSN: 1572-9613. DOI: 10.1007/s10955-016-1446-7.

Katzfuss, M. et al. (Oct. 1, 2016). "Understanding the Ensemble Kalman Filter." In: *The American Statistician* 70.4, pp. 350–357. ISSN: 0003-1305, 1537-2731. DOI: 10.1080/00031305.2016.1141709.

Kermack, W. O. et al. (Aug. 1927). "A Contribution to the Mathematical Theory of Epidemics." In: *Proc. R. Soc. Lond. A* 115.772, pp. 700–721. ISSN: 0950-1207, 2053-9150. DOI: 10.1098/rspa.1927.0118.

Khailaie, S. et al. (Jan. 28, 2021). "Development of the Reproduction Number from Coronavirus SARS-CoV-2 Case Data in Germany and Implications for Political Measures." In: *BMC Medicine* 19.1, p. 32. ISSN: 1741-7015. DOI: 10.1186/s12916-020-01884-4.

Khazaei, Y. et al. (Nov. 2, 2023). "Using a Bayesian Hierarchical Approach to Study the Association between Non-Pharmaceutical Interventions and the Spread of Covid-19 in Germany." In: *Sci Rep* 13.1, p. 18900. ISSN: 2045-2322. DOI: 10.1038/s41598-023-45950-2.

Kong, A. (1992). "A Note on Importance Sampling Using Standardized Weights." In: *University of Chicago, Dept. of Statistics, Tech. Rep* 348, p. 14.

Kong, A. et al. (Mar. 1994). "Sequential Imputations and Bayesian Missing Data Problems." In: *Journal of the American Statistical Association* 89.425, pp. 278–288. ISSN: 0162-1459. DOI: 10.1080/01621459.1994.10476469.

Koopman, S. J. et al. (1992). "Exact Score for Time Series Models in State Space Form." In: *Biometrika* 79.4, pp. 823–826. ISSN: 0006-3444. DOI: 10.2307/2337237. JSTOR: 2337237.

Koopman, S. J. et al. (2019). "Modified Efficient Importance Sampling for Partially Non-Gaussian State Space Models." In: *Statistica Neerlandica* 73.1, pp. 44–62. ISSN: 1467-9574. DOI: 10.1111/stan.12128.

Kraemer, M. U. G. et al. (May 2020). "The Effect of Human Mobility and Control Measures on the COVID-19 Epidemic in China." In: *Science* 368.6490, pp. 493–497. DOI: 10.1126/science.abb4218.

Laplace, P. S. (Aug. 1986). "Memoir on the Probability of the Causes of Events." In: *Statistical Science* 1.3, pp. 364–378. ISSN: 0883-4237, 2168-8745. DOI: 10.1214/ss/1177013621.

Lauer, S. A. et al. (2020). "The Incubation Period of Coronavirus Disease 2019 (COVID-19) from Publicly Reported Confirmed Cases: Estimation and Application." In: *Annals of internal medicine* 172.9, pp. 577–582. DOI: 10.7326/M20-0504.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford Statistical Science Series 17. Oxford : New York: Clarendon Press ; Oxford University Press. 298 pp. ISBN: 978-0-19-852219-5.

Lawless, J. F. (1994). "Adjustments for Reporting Delays and the Prediction of Occurred but Not Reported Events." In: *Canadian Journal of Statistics* 22.1, pp. 15–31. ISSN: 1708-945X. DOI: 10.2307/3315826.n1.

Liang, K.-Y. et al. (May 1995). "Inference Based on Estimating Functions in the Presence of Nuisance Parameters." In: *Statistical Science* 10.2, pp. 158–173. ISSN: 0883-4237, 2168-8745. DOI: 10.1214/ss/1177010028.

Lloyd-Smith, J. O. et al. (Nov. 2005). "Superspreading and the Effect of Individual Variation on Disease Emergence." In: *Nature* 438.7066, pp. 355–359. ISSN: 1476-4687. DOI: 10.1038/nature04153.

McGough, S. F. et al. (2020). "Nowcasting by Bayesian Smoothing: A Flexible, Generalizable Model for Real-Time Epidemic Tracking." In: *PLOS Computational Biology*. DOI: 10.1371/journal.pcbi.1007735. pmid: 32251464.

Midthune, D. N. et al. (Mar. 1, 2005). "Modeling Reporting Delays and Reporting Corrections in Cancer Registry Data." In: *Journal of the American Statistical Association* 100.469, pp. 61–70. ISSN: 0162-1459. DOI: 10.1198/016214504000001899.

Morf, M. et al. (Aug. 1975). "Square-Root Algorithms for Least-Squares Estimation." In: *IEEE Transactions on Automatic Control* 20.4, pp. 487–497. ISSN: 1558-2523. DOI: 10.1109/TAC.1975.1100994.

Mossong, J. et al. (Mar. 25, 2008). "Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases." In: *PLOS Medicine* 5.3, e74. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.0050074.

Nocedal, J. et al. (2006). *Numerical Optimization*. Second edition. Springer Series in Operation Research and Financial Engineering. New York, NY: Springer. 664 pp. ISBN: 978-0-387-30303-1 978-1-4939-3711-0.

Noufaily, A. et al. (2015). "Modelling Reporting Delays for Outbreak Detection in Infectious Disease Data." In: *Journal of The Royal Statistical Society Series A-statistics in Society*. DOI: 10.1111/rssa.12055.

Nowcasts Der COVID-19 Hospitalisierungsinzidenz (2022). URL: https://covid19nowcasthub.de/ (visited on 11/09/2022).

Rao, C. R. (2002). *Linear Statistical Inference and Its Applications*. 2. ed., Paperback ed. Wiley Series in Probability and Statistics. New York: Wiley. 625 pp. ISBN: 978-0-471-21875-3.

Ray, E. L. et al. (Aug. 22, 2020). "Ensemble Forecasts of Coronavirus Disease 2019 (COVID-19) in the U.S." In: *medRxiv*, p. 2020.08.19.20177493. DOI: 10.1101/2020.08.19.20177493.

Renshaw, A. E. et al. (1998). "A Stochastic Model Underlying the Chain-Ladder Technique." In: *British Actuarial Journal* 4.4, pp. 903–923. DOI: 10.1017/S1357321700000222.

Richard, J.-F. et al. (Dec. 1, 2007). "Efficient High-Dimensional Importance Sampling." In: *Journal of Econometrics* 141.2, pp. 1385–1411. ISSN: 0304-4076. DOI: 10.1016/j.jeconom.2007.02.007.

Ripley, B. D. (2009). *Stochastic Simulation*. Vol. 316. John Wiley & Sons.

Robert Koch-Institut (Oct. 1, 2021). *COVID-19-Hospitalisierungen in Deutschland*. Version 2021-10-01. Zenodo. DOI: 10.5281/ZENODO.5519056.

– (Feb. 7, 2022a). *SARS-CoV-2 Infektionen in Deutschland*. Version 2022-02-07. Zenodo. DOI: 10.5281/ZENODO.4681153.

– (Jan. 2022b). *Wöchentlicher Lagebericht Des RKI Zur Coronavirus-Krankheit-2019 (COVID-19)*.

Ross, N. (Jan. 1, 2011). "Fundamentals of Stein's Method." In: *Probab. Surveys* 8 (none). ISSN: 1549-5787. DOI: 10.1214/11-PS182.

Rubinstein, R. Y. (Sept. 1, 1999). "The Cross-Entropy Method for Combinatorial and Continuous Optimization." In: *Methodology and Computing in Applied Probability* 1.2, pp. 127–190. ISSN: 1573-7713. DOI: 10.1023/A:1010091220143.

Rubinstein, R. Y. et al. (2004). *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. New York, NY: Springer New York. ISBN: 978-1-4757-4321-0.

Rubinstein, R. Y. et al. (Nov. 6, 2009). "How to Deal with the Curse of Dimensionality of Likelihood Ratios in Monte Carlo Simulation." In: *Stochastic Models* 25.4, pp. 547–568. ISSN: 1532-6349. DOI: 10.1080/15326340903291248.

Salmon, M. et al. (2015). "Bayesian Outbreak Detection in the Presence of Reporting Delays." In: *Biometrical Journal*. DOI: 10.1002/bimj.201400159. pmid: 26250543.

Salzberger, B. et al. (Apr. 1, 2021). "Epidemiology of SARS-CoV-2." In: *Infection* 49.2, pp. 233–239. ISSN: 1439-0973. DOI: 10.1007/s15010-020-01531-3.

Schäfer, F. et al. (Jan. 2021). "Sparse Cholesky Factorization by Kullback–Leibler Minimization." In: *SIAM J. Sci. Comput.* 43.3, A2019–A2046. ISSN: 1064-8275. DOI: 10.1137/20M1336254.

Schlosser, F. et al. (Dec. 29, 2020). "COVID-19 Lockdown Induces Disease-Mitigating Structural Changes in Mobility Networks." In: *Proceedings of the National Academy of Sciences* 117.52, pp. 32883–32890. DOI: 10.1073/pnas.2012326117.

Schneble, M. et al. (Mar. 2021). "Nowcasting Fatal COVID-19 Infections on a Regional Level in Germany." In: *Biometrical Journal* 63.3, pp. 471–489. ISSN: 0323-3847, 1521-4036. DOI: 10.1002/bimj.202000143.

Schneider, W. (1986). *Der Kalmanfilter Als Instrument Zur Diagnose Und Schätzung Variabler Parameter in Ökonometrischen Modellen*. Arbeiten Zur Angewandten Statistik Bd. 27. Heidelberg: Physica-Verlag. 490 pp. ISBN: 978-3-7908-0359-4.

Shephard, N. (1994). "Partial Non-Gaussian State Space." In: *Biometrika* 81.1, pp. 115–131. DOI: 10.1093/biomet/81.1.115.

Shephard, N. et al. (Sept. 1, 1997). "Likelihood Analysis of Non-Gaussian Measurement Time Series." In: *Biometrika* 84.3, pp. 653–667. ISSN: 0006-3444. DOI: 10.1093/biomet/84.3.653.

Sherratt, K. et al. (June 16, 2022). *Predictive Performance of Multi-Model Ensemble Forecasts of COVID-19 across European Nations*. DOI: 10.1101/2022.06.16.22276024. Pre-published.

Song, J. et al. (May 1, 2021). "Maximum Likelihood-Based Extended Kalman Filter for COVID-19 Prediction." In: *Chaos, Solitons & Fractals* 146, p. 110922. ISSN: 0960-0779. DOI: 10.1016/j.chaos.2021.110922.

Stark, P. B. et al. (1995). "Bounded-Variable Least-Squares: An Algorithm and Applications." In: *Computational Statistics* 10, pp. 129–129. URL: https://digitalassets.lib.berkeley.edu/sdtr/ucb/text/394.pdf (visited on 06/05/2024).

Tierney, L. et al. (Mar. 1986). "Accurate Approximations for Posterior Moments and Marginal Densities." In: *Journal of the American Statistical Association* 81.393, pp. 82–86. ISSN: 0162-1459, 1537-274X. DOI: 10.1080/01621459.1986.10478240.

Tierney, L. et al. (Sept. 1989). "Fully Exponential Laplace Approximations to Expectations and Variances of Nonpositive Functions." In: *Journal of the American Statistical Association* 84.407, pp. 710–716. ISSN: 0162-1459, 1537-274X. DOI: 10.1080/01621459.1989.10478824.

Tomori, D. V. et al. (Mar. 26, 2021). "Individual Social Contact Data Reflected SARS-CoV-2 Transmission Dynamics during the First Wave in Germany Better than Population Mobility Data – an Analysis Based on the COVIMOD Study." In: p. 2021.03.24.21254194. DOI: 10.1101/2021.03.24.21254194.

Van de Kassteele, J. et al. (2019). "Nowcasting the Number of New Symptomatic Cases during Infectious Disease Outbreaks Using Constrained P-Spline Smoothing." In: *Epidemiology*. DOI: 10.1097/ede.0000000000001050. pmid: 31205290.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Wallinga, J. et al. (Feb. 22, 2007). "How Generation Intervals Shape the Relationship between Growth Rates and Reproductive Numbers." In: *Proc. R. Soc. B.* 274.1609, pp. 599–604. ISSN: 0962-8452, 1471-2954. DOI: 10.1098/rspb.2006.3754.

Wallinga, J. et al. (Sept. 15, 2004). "Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures." In: *American Journal of Epidemiology* 160.6, pp. 509–516. ISSN: 0002-9262. DOI: 10.1093/aje/kwh255.

White, H. (1982). "Maximum Likelihood Estimation of Misspecified Models." In: *Econometrica* 50.1, pp. 1–25. ISSN: 0012-9682. DOI: 10.2307/1912526. JSTOR: 1912526.

Whitt, W. (Nov. 1976). "Bivariate Distributions with Given Marginals." In: *The Annals of Statistics* 4.6, pp. 1280–1289. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/aos/1176343660.

Zeger, S. L. et al. (1989). "Statistical Methods for Monitoring the AIDS Epidemic." In: *Statistics in Medicine* 8.1, pp. 3–21. DOI: 10.1002/sim.4780080104.

Zhang, W. et al. (Jan. 2014). "Applications of the Cross-Entropy Method to Importance Sampling and Optimal Control of Diffusions." In: *SIAM J. Sci. Comput.* 36.6, A2654–A2672. ISSN: 1064-8275. DOI: 10.1137/14096493X.

Zhu, X. et al. (Oct. 1, 2021). "Extended Kalman Filter Based on Stochastic Epidemiological Model for COVID-19 Modelling." In: *Computers in Biology and Medicine* 137, p. 104810. ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2021.104810.

Abbreviations

BLUP best linear unbiased predictor. 42

CE-method Cross-Entropy method. vii, 2, 11, 29, 33–39, 41–44, 46–58, 66–75, 77
CLT central limit theorem. 39, 40, 45, 47
COD curse of dimensionality. 36
COVID-19 Coronavirus disease 2019. 1, 2, 4–6, 8, 18, 79, 80
CRN common random number. 36, 57, 58, 61, 64, 65

EF efficiency factor. 30, 31
EGSSM Exponential Family Partially Gaussian state space model. 23
EIS Efficient Importance Sampling. vii, 2, 11, 20, 29, 33, 42, 43, 45–50, 52, 57, 62, 65–75, 77
EKF Extended Kalman filter. 18
EnKF Ensemble Kalman filter. 18
ESS effective sample size. 29–31, 33, 48, 69

FFBS Forwards Filter, Backwards Sampling. 19, 63

GLSSM Gaussian linear state space model. vii, 11, 13–20, 48–50, 52, 57, 58, 60–62, 64, 66, 74

KL-divergence Kullback Leibler divergence. 27–29, 33, 34, 42, 55

LA Laplace approximation. vii, 23, 33, 48–50, 57, 64–68, 73, 74, 77
LCSSM Logconcave state space model. 13, 23, 49, 50, 57, 59, 67, 68

MC-integration Monte-Carlo integration. 19
MCMC Markov chain Monte Carlo. 14, 24, 37, 61
MLE maximum likelihood estimator. 13, 14, 17, 61, 64
MSE mean-squared error. 27, 30

NPI non-pharmaceutical intervention. 4, 5, 7

PGSSM Partially Gaussian state space model. 20, 21, 23, 26, 61, 62, 64, 65, 80
PSD positive semi-definite. 17

RKI Robert Koch-Institut. 1, 5, 81, 83, 88

SARS-CoV-2 Severe acute respiratory syndrome coronavirus 2. 4

SMC sequential Monte Carlo. 14, 24, 25, 61

SPD symmetric positive definite. 19, 30, 68, 73

SSM state space model. 2, 12–14, 20, 21, 24, 32, 33, 46, 48, 50–52, 60, 61, 73, 82

UKF unscented Kalman filter. 18

VM-method Variance-Minimization method. 47, 48

Declaration

Put your declaration here.

Ilmenau, October 2023