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. . .

iv

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This thesis consists of mostly unpublished work.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.

[1] J. Bracher et al. "A Pre-Registered Short-Term Forecasting Study of COVID-19 in Germany and Poland during the Second Wave." In: *Nature Communications* 12.1 (1 Aug. 27, 2021), p. 5173. ISSN: 2041-1723. DOI: 10.1038/s41467-021-25207-0. URL: https://www.nature.com/articles/s41467-021-25207-0 (visited on 09/30/2021).

[2] Johannes Bracher et al. "National and Subnational Short-Term Forecasting of COVID-19 in Germany and Poland during Early 2021." In: *Communications Medicine* 2.1 (1 Oct. 31, 2022), pp. 1–17. ISSN: 2730-664X. DOI: 10.1038/s43856-022-00191-8. URL: https://www.nature.com/articles/s43856-022-00191-8 (visited on 11/16/2022).

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

[4] Sara Grundel et al. "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 (Apr. 2022), S145–S169. ISSN: 0363-0129, 1095-7138. DOI: 10.1137/20M1377783. URL: https://epubs.siam.org/doi/10.1137/20M1377783 (visited on 11/16/2022).

[5] Sara M. Grundel et al. "How to Coordinate Vaccination and Social Distancing to Mitigate SARS-CoV-2 Outbreaks." In: *SIAM Journal on Applied Dynamical Systems* 20.2 (Jan. 1, 2021), pp. 1135–1157. DOI: 10.1137/20M1387687. URL: https://epubs.siam.org/doi/abs/10.1137/20M1387687 (visited on 01/21/2022).

[6] Thomas Hotz et al. "Monitoring the Spread of COVID-19 by Estimating Reproduction Numbers over Time." Apr. 18, 2020. arXiv: 2004.08557 [q-bio, stat]. URL: http://arxiv.org/abs/2004.08557 (visited on 07/20/2020).

[7] K. Sherratt et al. *Predictive Performance of Multi-Model Ensemble Forecasts of COVID-19 across European Nations.* June 16, 2022. DOI: 10.1101/2022.06.16.22276024. URL: http://medrxiv.org/lookup/doi/10.1101/2022.06.16.22276024 (visited on 11/28/2022). preprint.

Du musst bereit sein Dinge zu tun.

— A meme on the internet, 2022.

Acknowledgments

Put your acknowledgments here.

Contents

Chapter 1

Introduction

Chapter 2

Epidemiological considerations

Contributions of this chapter

- ...

- COVID-19 induced unprecedented interest in epidemiological modelling from all disciplines, but also mathematics

- this chapter highlights challenges that epi modelling brings and what desirable outcomes would be from an applied perspective

- mathematical epidemiology concerns itself with modelling epidemiolocial systems, from small (local outbreaks) to large (epi/pandemics)

- conclusions from analysis only as good as the model and the data are

- dependening on goal and circumstances different methods are applicable

- by its nature, data are observational so causal claims difficult

- in this thesis I focus on models for larger-scale epidemics, techniques would be flexible enough to deal with smaller scale as well, as long as latent states are gaussian

2.1   Objectives of epidemiological modelling

Monitoring

- monitoring is real-time scenario, interested in current developments, i.e. recent past and near future. complicated by potentially slow reporting, data revisions

- informs decision makers on whether measures should be taken

- ForecastHub(s) provide platform that creates ensemble forecast to obtain better predictions [6, 7, 43, 57]

Retrospective Analysis

- evaluation of measures taken, want interpretation as causal as possible

- informs decision makes on which measures were effective and how much

- difficult due to usual reasons: poor data quality, observational data, causual structure difficult, early/late adoption makes timing of measurements difficult

- cite some papers that did this [8, 20, 32]

Scenario Modelling

- concerns itself with modelling the impact that variants, seasonality etc. have in specific scenarios

- find out whether there is already paper of ECDC to cite

2.2   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

2.3 Measures of epidemic spread

This section consists of the ideas published in [24], but has been rewritten to fit better into this thesis.

- not only epidemic spread but also speed of proliferation is of interest, enables forecasts
- measuring speed difficult: data problems ... (look at AK book article)
-

2.3.1 Growth Factor

2.3.2 Reproduction number

2.3.3 Other indicators

2.3.4 Usefulness of indicators

2.4 Dessiderata for epidemiological models

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

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 of this chapter

- asymptotics for CE/EIS
- CE method for SSMs
- comparison of CE and EIS, theoretically
- simulation studies for the comparison

State space models (SSMs) form a versatile class of statistical models which allow to model non-stationary time series data and come along with straight-forward interpretation. 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 theses states. By exploiting this structure, inference in SSMs becomes computationally efficient, i.e. the complexity of algorithms is linear in the number n of time points considered.

An additional advantage, that will become more explicit in Section 3.1, is that SSMs allow to interpret the modeled dynamics of latent states which makes

Definition 3.1 (State Space Model). `fix n, n-1`

A **SSM** is a discrete time stochastic process $(X_t, Y_t)_{t=0,\ldots,n-1}$ taking values in a measurable space $\mathcal{X} \times \mathcal{Y}$ such that

1. The marginal distribution of the **states** $(X_0, \ldots, X_{n-1})$ is a discrete time Markov process, i.e. for $t = 1, \ldots, n-1$

$$\mathbf{P}\left(X_t \in B | X_0, \ldots, X_{t-1}\right) = \mathbf{P}\left(X_t \in B | X_{t-1}\right) \tag{3.1}$$

   for all measurable $B \subseteq \mathcal{X}$.

2. 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.

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

   for all measurable $B \subseteq \mathcal{Y}$.

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

`picture of dependency structure`

Remark 3.1. Contrary to the standard definition of a SSM, our Definition 3.1 allows Y_t to depend on Y_{t-1}. 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, most computations and inferences in this thesis will be conditioned on a single set of observations Y. We thus adopt a Bayesian perspective and as such Y may be treated as a fixed and Y_t only depends on X_t. We will not be interested in frequentist properties of the models concerning repeated sampling of Y. The only exception to this is in simulation studies where we sample from the joint distribution of (X, Y).

`do we really need this here?`

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.

In most models 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 Borel σ-Algebras.

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

`check whether there are models that violate this`

Notation 3.1 (Densities, conditional densities). I 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 stem from a parametric model parametrized by $\theta \in \Theta \subseteq \mathbf{R}^k$ 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

$$p_\theta(x, y) = p_\theta(x_0, \ldots, x_{n-1}, y_0, \ldots, y_{n-1})$$
$$= p_\theta(x_0) \prod_{t=1}^{n-1} p_\theta(x_t|x_{t-1}) \prod_{t=0}^{n-1} p_\theta(y_t|x_t, y_{t-1}),$$

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.

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

1. Choosing a suitable, usually parametric, class of SSMs that include the effects of interest.

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

3. 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 1, 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 behaviour 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, leading to the class of logconcave Gaussian state space models (Section 3.3).

Regarding the second step, Item 2, a frequentist practitioner will want to perform maximum likelihood inference on θ. While asymptotic confidence intervals for θ can be derived both theoretically and practically [14, Chapter 7], they are, in the context of this thesis, usually of little interest. We choose to view this fitting as an Empirical Bayes procedure and our main practial interest lies in analyzing the posterior distribution $X|Y$.

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

$$p(y) = \int_{\mathcal{X}^n} p(x, y) \, \mathrm{d}x, \tag{3.2}$$

which is usually not analytically available. Direct numerical evaluation of Equation (3.2) 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), an alternative would be particle filters [12].

The performance of these simulations depends crucially on constructing distributions that are close to the posterior $p(x|y)$ but are easy to sample from. To this end, we construct suitable Gaussian state space models (Section 3.5) in which sampling from the posterior is analytically possible. This will be a good strategy if the target posterior $p(x|y)$ can be well approximated by a Gaussian distribution — otherwise, we may want to account for multiple modes by considering mixtures of Gaussian state space models or account for heavy tails with t-distributed errors (Section 3.6).

3.1   Modelling epidemiological dessiderata with state space models

3.2   Linear Gaussian state space models

GLSSM 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 GLSSM is a state space model where states obey the transition equation

$$X_{t+1} = A_t X_t + u_t + \varepsilon_{t+1} \qquad\qquad t = 0, \ldots, n-1, \tag{3.3}$$

and observations obey the observation equation

$$Y_t = B_t X_t + v_t + \eta_t \qquad\qquad t = 0, \ldots, n. \tag{3.4}$$

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** ε_{t+1} and **measurement noise** η_t are independent from one another and from the starting value $X_0 \sim \mathcal{N}(\mathbf{E}X_0, \Sigma_0)$. 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 $u_t \in \mathbf{R}^m, t = 0, \ldots, n-1$, $v_t \in \mathbf{R}^p, t = 0, \ldots, n$ are deterministic biases.

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 $(X_0, \varepsilon_1, \ldots, \varepsilon_n, \eta_0, \ldots, \eta_n)$.

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

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

for non-singular Σ_{YY}.

Then $X|Y = y$ is also a Gaussian distribution with conditional expectation

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

and conditional covariance matrix

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

In particular, if $Y = BX + \varepsilon$ for a matrix $B \in \mathbf{R}^{p \times m}$ and $\exists \varepsilon \sim \mathcal{N}(0, \Omega)$ for $\Omega \in \mathbf{R}^{p \times p}$ independent of X, then, as $\mu_Y = B\mu_X$, $\Sigma_{XY} = \Sigma_{YX}^T = \Sigma_{XX}B^T$ and $\Sigma_{YY} = B\Sigma_{XX}B^T + \Omega$, we have

$$\mu_{X|Y=y} = \mu_X + K(y - B\mu_X)$$

and

$$\Sigma_{X|Y=y} = \Sigma_{XX} - K\Sigma_{YY}K^T = (I - KB)\Sigma_{XX}$$

with $K = \Sigma_{XX}B^T \left(B\Sigma_{XX}B^T + \Omega\right)^{-1}$.

Proof. For the first statement, we refer the reader to [14, Chapter 4, Lemma 1]. The second statement follows from substituting the value of K.

$\square$

Let us denote by $\hat{X}_{t|s}$ the conditional mean of X_t given observations $Y_{:s}$ and by $\Xi_{t|s}$ the conditional covariance matrix of X_t given $Y_{:s}$. Then $X_t|Y_{:s} \sim \mathcal{N}\left(\hat{X}_{t|s}, \Xi_{t|s}\right)$. For a given t, three values of s are of 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 $\hat{Y}_{t|s} = \mathbf{E}\left(Y_t|Y_{:s}\right)$ to be the conditional expectation of Y_t given $Y_{:s}$, note that $\hat{Y}_{t|s} = Y_t$ if $s \geq t$. Finally, let $\Psi_{t|s} = \mathrm{Cov}\left(Y_t|Y_{:s}\right)$ be the conditional covariance matrix of Y_t given $Y_{:s}$. Again $\Psi_{t|s} = 0$ if $s \geq t$.

As the joint distribution of states is Gaussian, these conditional expectations are given by the best linear unbiased predictor (BLUP) of X_t $Y_{:s}$.

> cite

> Note that the filtering distributions does not specify a valid joint distribution for the states, but the smoothing does.

These distributions may be obtained efficiently using the celebrated Kalman filter and smoother algorithms

> cite correclty

, which we state here for completeness.

Algorithm 1 Kalman filter

Require: GLSSM (Definition 3.2), observations $Y_0, \ldots, Y_n$.

$A_{-1} \leftarrow I \in \mathbf{R}^{m \times m}$ ▷ Identity Matrix
$u_{-1} \leftarrow \mathbf{0} \in \mathbf{R}^m$
$\hat{X}_{-1|-1} \leftarrow \mathbf{E}X_0$
$\Xi_{0|-1} \leftarrow \Sigma_0$

for $t \leftarrow 0, \ldots, n$ **do**

$\quad \hat{X}_{t|t-1} \leftarrow A_{t-1}\hat{X}_{t-1|t-1} + u_{t-1}$ ▷ prediction
$\quad \Xi_{t|t-1} \leftarrow A_{t-1}\Xi_{t-1|t-1}A_{t-1}^T + \Sigma_t$
$\quad \hat{Y}_{t|t-1} \leftarrow B_t\hat{X}_{t|t-1} + v_t$
$\quad \Psi_{t|t-1} \leftarrow B_t\Xi_{t|t-1}B_t^T + \Omega_t$

$\quad K_t \leftarrow \Xi_{t|t-1}B_t^T\Psi_{t|t-1}^{-1}$ ▷ filtering
$\quad \hat{X}_{t|t} \leftarrow \hat{X}_{t|t-1} + K_t(Y_t - \hat{Y}_{t|t-1})$
$\quad \Xi_{t|t} \leftarrow \Xi_{t|t-1} - K_t\Psi_{t|t-1}K_t^T$
end for

> make n, n-1 consistent (ISSSM uses from 0 to n)

In Algorithm 1 every time point $t = 0, \ldots, n$ 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 [14, Chapter 4].

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 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.

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 [40] may be used. It is based on Cholesky roots of the involved covariance matrices, ensuring them to be positive-semi definite. If ... one may use the information filter, that operates on the precision matrices instead of covariance matrices

write some more about them here

.

Algorithm 2 Kalman smoother

Require: todo

Notice that the Kalman filter calculates the likelihood $p(y)$ while filtering — this is possible because of the dependency structure of the state space model — this makes inference via maximum likelihood possible in GLSSMs.

To ensure numerical stability in these algorithms, the square root filter and smoother [40] may be used, see also [55] for an accessible introduction to it and other variants.

The Kalman smoother computes the marginal distributions $X_t|Y$ for $t = 0, \ldots, n-1$ and, owing to the Markov structure of the states, these are enough to specify the joint distribution $X|Y$, allowing to simulate from it.

Algorithm 3 Forwards filter, backwards smoother [21, Proposition 1]

Require: TODO

The modeling capacity of GLSSM is, however, limited: most interesting phenomena follow neither linear dynamics nor are well modeled by a Gaussian distribution. 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   Logconcave Gaussian state space models

- replace gaussian observations with log concave observations
- motivation for logconcave distributions: posterior has unique mode, because up to constants $\log p(x|y) = -\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu) + \log p(y|x)$ so $\log p(x|y)$ is concave
- not restricted to same type of distribution per time step (though in ISSSM it will be)
- Laplace approximation sensible for these types of models: single mode
- special case: exponential family distributions

The distribution of observations is never Gaussian - all statisticians may hope for is that the data-generating mechanism is close enough to a Gaussian distribution that inferences made carry over. For epidemiological models, Gaussian distributions are 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

auto/Cref?

. Both the Poisson and negative binomial belong to the class of exponential family distributions. As such, their densities have a simple structure, allowing only for a linear interaction between the

natural parameter and the densities argument. We refer to [9] for a comprehensive treatment of exponential families and use their definitions throughout this chapter.

Definition 3.3 (exponential family)**.** Let μ be a σ-finite measure on $\mathbf{R}^p$ and denote by

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

the set of parameters θ such that the moment-generating function of μ is finite. For every $\theta \in \Theta$

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

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

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

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

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

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

Conversely, let $\mathbf{P}_\theta, \theta \in \Theta$ 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 exists a reparametrization $\eta : \Theta \to \mathbf{R}^p$, a statistic $T : \mathcal{Y} \to \mathbf{R}^p$ and functions $Z : \Theta \to \mathbf{R}$, $h : \mathcal{Y} \to \mathbf{R}$ exist, such that

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

then we call $\mathbf{P}_\theta, \theta \in \Theta$ and $p_\theta, \theta \in \Theta$ a p**-dimensional exponential family**. If $\eta(\theta) = \theta$ we call θ the canonical parameter. If $T(y) = y$, we call y the canonical observation. By reparametrization (in θ) and sufficiency (in y) every p-dimensional exponential family can be written as an equivalent standard exponential family, see the elaborations in [9, Chapter 1].

> necessary?

Definition 3.4 (curved exponential family)**.**

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

> later section

.

Lemma 3.2 (log-concavity of exponential family distributions)**.** *Let* $p_\theta, \theta \in \Theta$ *be a natural dimensional exponential family and* Θ *open in* $\mathbf{R}^p$*. In this case* $\theta \to \log p_\theta(y)$ *is concave for every* $y \in \mathbf{R}^p$*.*

Proof. As $\log p_\theta(y) = -\log Z(\theta) + \theta^T y$ it suffices to show that $\log Z(\theta)$ is convex. However, $\log Z(\theta)$ is the cumulant generating function of the base measure μ which is known to be convex

> more reasoning? differentiate under integral, check if dominated convergence applies, or look for a ref

. $\qquad \square$

We now generalize Definition 3.2 to allow for non-gaussian observations by replacing the observation equation Equation (3.4) by more general exponential families.

Definition 3.5 (Logconcave state space model (LCSSM)). A **Logconcave state space model (LCSSM)** is a SSM where states obey the transition equation

$$X_{t+1} = A_t X_t + u_t + \varepsilon_{t+1}$$

and 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, additionally, matrices $B_t \in \mathbf{R}^{p \times m}$ exist, such that for the signal $s_t = B_t x_t \in \mathbf{R}^p$ 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 $h_t^i : \mathbf{R} \to \mathbf{R}, Z_t^i : \mathbf{R} \to \mathbf{R}, \eta_t^i : \mathbf{R} \to \mathbf{R}, T : \mathbf{R} \to \mathbf{R}, i = 1, \ldots p$, we say the Logconcave state space model (LCSSM) has a **linear signal**.

> LCSSM only logconcave observations would suffice, then EF is just an instance of this

Remark 3.2. 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 relevant t.

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) (see **??**), or simulation based inference, e.g. importance sampling or MCMC-methods, are used. Similarly, fitting hyperparameters θ by maximum likelihood inference becomes more difficult as evaluating $\ell(\theta) = p(y) = \int p(x, y) \, dx$ is not analyically avaiable, thus requiring numerical or simulation methods for evaluation and gradient descent or EM-techniques for optimization (

> really EM?

.

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

> move this to the LA section

Under appropriate assumptions on the observation distributions, the LA is given by a (possibly degenerate) GLSSM [15, 56], though the derivations and computations are more involved in the nonlinear case or if the observation densities are not log-concave.

Theorem 3.1 (LA of LCSSM).

> 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 [4, 49]. However, incidences observed from the SARS-CoV-2 epidemic tend to follow a negative binomial distribution [10].

3.4 Importance Sampling

> some intro

Importance sampling is a simulation technique that allows to approximate expectations by sampling from a tractable approximation, the proposal, to the measure of interest, the target, and weighting 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 allows to perform Monte-Carlo integration even if sampling from the target is not possible, and second it allows 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 sequential Monte Carlo (SMC). Recall from Chapter

> insert

that this inference concerns three objectives: maximum likelihood estimation, i.e. evaluation and optimization of the likelihood, 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

> cite brook handbook

is a simulation technique that allows to generate correlated samples from a distribution by constructing a Markov chain that has as its invariant distribution the desired distribution. In the most general method, 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 a lot of current research in MCMC methods deals with this

> quotes

curse of dimensionality

> cite something

.

> dis-advantages MCMC over IS

SMC [12], sometimes called a particle filter, uses 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. [12, Chapter 14.3] 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. Under j stochastic gradient ascent

> more

This chapter proceeds with a general treatment of importance sampling. Subsequently, we will focus our attention on LCSSMs and how we can exploit their structure to perform importance sampling efficiently.

Suppose we have a function $h : \mathcal{X} \to \mathbf{R}$ whose integral

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

we want to compute. Furthermore, suppose that we can write

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

for a probability measure $\mathbf{P}$ and function $f : \mathcal{X} \to \mathbf{R}$. Let $\mathbf{G}$ be another 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

$$\zeta = \int_{\mathcal{X}} h(x) \, \mathrm{d}x = \int_{\mathcal{X}} f(x) \, \mathrm{d}\mathbf{P}(x) = \int_{\mathcal{X}} v(x) \, \mathrm{d}\mathbf{G}(x)$$

which suggests to estimate ζ by Monte-Carlo integration:

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

for $X_i \overset{\text{i.i.d}}{\sim} \mathbf{G}$, $i = 1, \ldots, N$. Here we call $\hat{\zeta}$ the importance sampling estimate of ζ.

A classical result is that the minimum MSE proposal $\mathbf{G}^*$ has a closed form, which can be shown by a simple application of Jensen's inequality.

Proposition 3.1 ([12, 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'd have to obtain $\mathbf{P}(f)$, the overall target of our endeavor. Additionally, sampling from $\mathbf{G}^*$ is not guaranteed to be practically feasible.

If one is not interested in a particular h but rather in an approximation of $\mathbf{P}$, and $\mathbf{P}$ is absolutely continuous with respect to $\mathbf{G}$, then one may view

$$\hat{\mathbf{P}}_N = \frac{1}{N} \sum_{i=1}^{N} v(X_i)\delta_{X_i} \tag{3.5}$$

as a particle approximation of $\mathbf{P}$, in the sense that for sufficiently well behaved test functions f, $\mathbf{P}(f) \approx \hat{\mathbf{P}}_N(f)$. In this setting [1] shows that the random measure $\hat{\mathbf{P}}_N$ converges to $\mathbf{P}$ at rate $\mathcal{O}\left(\frac{1}{N}\right)$

check

in an appropriate metric.

To perform importance sampling one must be able to evaluate the weights v. In a Bayesian setting this is usually infeasible: if $\mathbf{P}$ is a posterior distribution then the integration constant of its density is intractable. In this case one can usually evaluate the weights up to a constant, i.e. $w(x) \propto_x \frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{G}}(x)$ is available. The missing constant is then $\int w(x) \, \mathrm{d}\mathbf{G}$ which is itself amenable to importance sampling. This leads to the self-normalized importance sampling weights $W_i = \frac{w(X_i)}{\sum_{i=1}^{N} w(X_i)}$ and Monte Carlo estimates $\hat{\zeta} = \sum_{i=1}^{N} W_i f(X_i)$ and particle approximation $\hat{\mathbf{P}}_N = \sum_{i=1}^{N} W_i \delta_{X_i}$.

In both cases one can show that once the second moment of w with respect to $\mathbf{G}$ exists the Monte-Carlo estimates are consistent and asymptotically normal at the usual rates, see [12, Chapter 8]. However, the finite sample variance of $\hat{\zeta}$, and thus the practical performance of the procedure, depends on the variance of $w \cdot f$ under $\mathbf{G}$, and thus on the proposal $\mathbf{G}$. [1] show that the expected mean squared TV distance

`check if true`

between $\hat{\mathbf{P}}_N$ and $\mathbf{P}$ may be bounded, up to a constant, by the second moments of w

`more extensive discussion of ρ`

. In addition, they provide bounds that involve the KL-divergence

$$\mathcal{D}_{\mathrm{KL}}\left(\mathbf{P}||\mathbf{G}\right) = \int \log \frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{G}} \, \mathrm{d}\mathbf{P},$$

fostering the intuition that $\mathbf{G}$ should be close to $\mathbf{P}$ for the particle approximation $\hat{\mathbf{P}}_N$ to be close to $\mathbf{P}$.

[11] provides the following theorem (using our notation), that relates the performance of importance sampling to both the Kullback Leibler divergence (KL-divergence) and the tail behavior of the log weights.

Theorem 3.2 ([11, Theorem 1.1]). *If $N = \exp\left(\mathcal{D}_{KL}\left(\mathbf{P}||\mathbf{G}\right) + t\right)$ for $t \geq 0$, and $f \in L^2(\mathbf{P})$ then*

$$\mathbf{G}\left|\mathbf{P}_N f - \mathbf{P}f\right| \leq \|f\|_2 + 2\sqrt{\mathbf{P}\left(\log w > \mathcal{D}_{KL}\left(\mathbf{P}||\mathbf{G}\right) + \frac{t}{2}\right)}.$$

`really P in the log weights?`

[11, Theorem 1.2] provides a similar result for autonormalised importance sampling.

`more referenceable defs in this chapter`

To judge the convergence of importance sampling several criteria are discussed in the literature. The classic effective sample size (ESS)[33]

$$\mathrm{ESS} = \frac{1}{\sum_{i=1}^N W_i^2}$$

arises from the asymptotic efficiency of importance sampling estimates and is easy to interpret, though care has to be taken in particular settings

`find literature`

. Assessing convergence through the variance of $\hat{\mathbf{P}}_N$ is, while natural, flawed [11] and should be avoided. As a remedy [11] suggest the heuristic $q_N = \mathbf{E}Q_N$ where

$$Q_N = \max_{1 \leq i \leq N} W_i.$$

This judges whether importance sampling has collapsed to just a few particles and is itself amenable to Monte-Carlo integration.

In the following sections, we will predominantly take the position that we are interested in the particle approximation $\hat{\mathbf{P}}_N$ of the form Equation (3.5) 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, most problems considered in this thesis exist in a Bayesian context where $\mathbf{P}$ is usually a posterior distribution, i.e. $\mathbf{P} = \mathbf{P}^{X|Y=y}$ for some random variables X and Y. Should the appropriate densities exist, evaluating the weights amounts to calculating

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

In these situations $p(y) = \int p(x, y) \mathrm{d}\mu(x)$ 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 $\mathbf{P}$. 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.

> discuss problems/remedies of high dimensional importance sampling

In the context of this thesis, importance sampling serves as the main tool to facilitate inference in SSMs, both for maximum likelihood estimation and access to the smoothing distribution. As the dimension of the state space is $n \cdot m^2$, which will usually be large, we have to perform importance sampling efficiently, exploiting the dependence structure.

> decide where to put this

As the likelihood of a general state space model is neither analytically nor numerically tractable one has to resort to Monte-Carlo techniques. Recall that the likelihood is a high-dimensional integral of the form

$$\ell(\theta) = p_\theta(y) = \int p_\theta(y, x)\, \mathrm{d}x = \int p_\theta(y|x)p_\theta(x)\, \mathrm{d}x = \mathbf{E}p_\theta(y|X).$$

By the standard law of large numbers we can approximate $\ell(\theta)$ by

$$\hat{\ell}(\theta) = \frac{1}{N} \sum_{i=1}^{N} p_\theta(y|X^i)$$

for $N \in \mathbf{N}$ samples $X^i \overset{\text{i.i.d}}{\sim} p(x)$. However, the variance of $\hat{\ell}(\theta)$ is likely to be very high if samples X^i are drawn from the prior distribution $p(x)$ as they are not informed by the observations y. As $p_\theta(x|y) \propto p_\theta(x, y)$ a more promising approach would be to use samples $X^i \sim p_\theta(x|y)$, but this distribution is usually not available.

While Bayesian computational approaches such as MCMC[**Brooks2011Handbook**] are able to generate (approximate) samples from this posterior distribution, importance sampling tries to find a distribution close to the target and re-weighs samples to ensure unbiased estimates of $\ell(\theta)$.

- importance sampling as a variance reduction technique
- importance sampling as a technique to make intractable distributions tractable
- importance sampling vs. other methods:
    - vs. ABC
    - vs. MCMC
    - vs. INLA (isn't this MCMC?)
- measuring how good IS performs: ESS and other measures
- related results regarding performance of IS (Chatterje, Agapiou)

3.4.1   Laplace approximation (LA)

Laplace approximation (LA) goes back to Laplace [35] who invented the technique to approximate moments of otherwise intractable distributions. Since [58, 59] rediscovered its use to approximate posterior means and variances, it has been a staple method for statisticians

> cite something

.

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 and the mode exists and is unique, we have

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

where H is the Hessian of $\log p$ evaluated at the mode. 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 $\Sigma = -H^{-1}$.

If p is log-concave in x, H is guaranteed to be negative semidefinite and the LA yields an actual Gaussian distribution.

> what to do if we cannot gurantee this

The main advantage of the LA is that 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 [15, 56], even in the degenerate case where H is indefinite [29].

> incorporate some of the citations from How good is your LA

However, as the LA is a local approximation, it may be an inappropriate description of the global behavior of the target, see Example 3.2 for a breakdown of LA and the simulation studies presented in Section 3.8. Additionally, even if 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 Cross-entropy method (CE-method)

To provide a global approximation to the target, the CE-method[50, 51] selects from a family $(\mathbf{G}_\psi)_{\psi \in \Psi}$ of proposals the one that minimizes the KL-divergence to the target. Thus, the CE-method finds ψ_{CE} which solves the following optimization problem

$$\psi_{\mathrm{CE}} = \mathrm{argmin}_{\psi \in \Psi} \, \mathcal{D}_{\mathrm{KL}}\left(\mathbf{P} || \mathbf{G}_\psi\right)$$
$$= \mathrm{argmin}_{\psi \in \Psi} \int \log \frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{G}_\psi} \, \mathrm{d}\mathbf{P}.$$

If $\mathbf{P}$ and $\mathbf{G}_\psi$ possess densities p and g_ψ w.r.t. some common measure μ, the same for all ψ, we may reformulate the optimization problem to maximize the cross-entropy between p and g_ψ instead:

$$\psi_{\mathrm{CE}} = \mathrm{argmin}_{\psi \in \Psi} \int p(x) \log p(x) \, \mathrm{d}\mu(x) - \int p(x) \log g_\psi \, \mathrm{d}\mu(x)$$
$$= \mathrm{argmax}_{\psi \in \Psi} \int p(x) \log g_\psi(x) \, \mathrm{d}\mu(x), . \tag{3.6}$$

Note that the first integral does not depend on ψ. The assumption of such a dominating measure is not restrictive: otherwise the KL-divergence is infinite.

An attractive property of the CE-method is that if $\mathbf{G}_\psi$ form an exponential family with natural parameter $\psi \in \mathbf{R}^p$, the optimal ψ_{CE} only depends on certain moments of $\mathbf{P}$. Indeed, for $\log g_\psi(x) = \log h(x) - \log Z(\psi) + \psi^T T(x)$ we have

$$\int p(x) \log g_\psi(x) \, \mathrm{d}\mu(x) = \mathbf{P} \log h - \log Z(\psi) + \psi^T \mathbf{P}T.$$

As $\log Z(\psi)$ is the cumulant-generating function of $\mathbf{G}_\psi$ it is, under appropriate regularity conditions, smooth. Thus the optimal ψ_{CE} solves

$$\mathbf{P}T = \nabla_\psi \log Z(\psi_{\mathrm{CE}}) = \mathbf{G}_{\psi_{\mathrm{CE}}}T,$$

and so reduces the task at hand to match the moments of the sufficient statistic of the target and proposal. Given $\mathbf{P}T$, this system of equations can, in many cases, be solved analytically or by gradient descent algorithms.

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 $X^1,\dots,X^N \overset{\mathrm{i.i.d}}{\sim} \mathbf{G}$ and auto-normalized importance sampling weights W^i and in turn estimate ψ_{CE} by $\hat{\psi}_{\mathrm{CE}}$ solving

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

Thus $\hat{\psi}_{\mathrm{CE}}$ is a Z-estimator, i.e. an estimator that arises from solving a random system of equations, and we can analyze its asymptotic behavior using standard results from the theory of Z-estimators. The following theorem of [61] will be useful in analyzing the asymptotic behavior of the estimators we consider in this thesis. We state it here, using our notation, for completeness.

Theorem 3.3 (asymptotic variance of Z-estimators, [61, Theorem 5.21]). *For every ψ in an open subset of $\mathbf{R}^k$, let $x \mapsto f_\psi(x)$ be a measurable vector-valued function such that, for every ψ_1 and ψ_2 in a neighborhood of ψ_0 and a measurable function $\dot{f}$ with $\mathbf{G}\dot{f} < \infty$,*

$$\|f_{\psi_1}(x) - f_{\psi_2}(x)\| \le \dot{f}(x)\|\psi_1 - \psi_2\|. \tag{LL}$$

Assume that $\mathbf{G}\|f_{\psi_0}\| < \infty$ and that the map $\psi \mapsto \mathbf{G}f_\psi$ is differentiable at ψ_0, with nonsingular derivative matrix B^{-1}. Let $X_1,\dots,X_N \overset{i.i.d}{\sim} \mathbf{G}$ and $\hat{\mathbf{G}}_N = \sum_{i=1}^N \delta_{X_i}$. If $\hat{\psi}_N$ fulfills $\hat{\mathbf{G}}_N f_{\hat{\psi}_N} = o_P\left(N^{-\frac{1}{2}}\right)$, and $\hat{\psi}_N \to \psi_0$ in probability, then

$$\sqrt{N}\left(\hat{\psi}_N - \psi_0\right) \overset{\mathcal{D}}{\to} \mathcal{N}(0, BMB^T), \tag{3.7}$$

where $M = \mathbf{G}f_{\psi_0}f_{\psi_0^T}$.

Notation 3.2 (central limit theorem for Z-estimators). The central limit theorems derived in this and the next section will make frequent use of Theorem 3.3. We will use the following consistent notation in the statement of theorems and their proofs:

- $f_\psi(x) : \mathbf{R}^k \to \mathbf{R}^k$ the estimating equation
- $B = (\mathbf{G}\partial_\psi f_\psi)^{-1}$ the bread matrix
- $M = \mathbf{G}f_\psi f_\psi^T$ the meat matrix
- $V = BMB$ the asymptotic covariance matrix
- $\log g_\psi(x) = \psi^T T(x) + \log h(x) - \log Z(\psi)$ the density of the natural exponential family considered
- $\dot{z}(\psi) = \nabla_\psi \log Z(\psi) = \mathbf{G}_\psi T$ the derivative of the log-normalizing constant $\psi \mapsto \log Z(\psi)$
- $\ddot{z}(\psi) = \partial_\psi \dot{z}(\psi) = \mathrm{Cov}_{\mathbf{G}_\psi}T$ the Hessian of the log-normalizing constant $\psi \mapsto \log Z(\psi)$

The naming of B and M stems from the sandwich estimator [62], where B is the Jacobian of the estimating equations $\mathbf{P}f_\psi = 0$ and M is the covariance matrix of f_ψ under $\mathbf{P}$.

Theorem 3.4 (consistency of $\hat{\psi}_{\mathrm{CE}}$). ⬛ *from vdV*

Theorem 3.5 (asymptotic normality of $\hat{\psi}_{\mathrm{CE}}$). ⬛ *require unique solution?*

Let $\mathbf{G}_\psi$ form a natural exponential family with densities $g_\psi(x) = \frac{h(x)}{Z(\psi)} \exp\left(\psi^T T(x)\right)$ w.r.t. μ. Let $\mathbf{G}, \mathbf{P}$ be two other probability measures such that $\mathbf{G} \ll \mathbf{P}$ and let $W = \frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{G}}$ be the normalized importance sampling weights. Suppose further that

(A1) $\mathbf{G}_{\hat{\psi}_{CE}}T = \hat{\mathbf{P}}_N T$ μ-a.s. has a unique solution $\hat{\psi}_{CE}$,

(A2) $\psi \mapsto \nabla_\psi \log Z(\psi)$ is locally Lipschitz around ψ_{CE},

(A3) W, T and WT possess finite second moments w.r.t. $\mathbf{G}$,

(A4) the Fisher information $I(\psi_{CE})$ is positive definite and equal to $-\ddot{z}(\psi_{CE})$, additionally $\psi \mapsto I(\psi)$ is continuous, and

(A5) regularity conditions

> choose correct ones to make $I(\psi) = -\ddot{z}(\psi)$

.

Then, as N goes to ∞,

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

where

$$V_{CE} = B_{CE} M_{CE} B_{CE},$$

with

$$B_{CE} = I(\psi_{CE})^{-1},$$
$$M_{CE} = Cov_{\mathbf{G}}\left(W(T - \mathbf{G}_{\psi_{CE}}T)\right).$$

Moreover $\mathbf{G}(W(T - \mathbf{G}_{\psi_{CE}}T))) = 0$, so we may estimate V_{CE} consistently by plug-in:

$$\hat{V}_{CE} = I(\hat{\psi}_{CE})^{-1}\left(\sum_{i=1}^{N} W_i^2 \left(T(X^i) - \mathbf{G}_{\hat{\psi}_{CE}}T\right)\left(T(X^i) - \mathbf{G}_{\hat{\psi}_{CE}}T\right)^T\right) I(\hat{\psi}_{CE})^{-1}.$$

Proof. We check that the conditions of the central limit theorem for Z-estimators (Theorem 3.3) are fulfilled. This proof uses the notation established in Notation 3.2. Consider the estimating equations for ψ_{CE}

$$x \mapsto f_\psi(x) = \nabla_\psi\left(w(x)\log g_\psi(x)\right) = w(x)T(x) - w(x)\dot{z}(\psi),$$

where $w(x)$ are the unnormalized importance sampling weights. By **(A1)** $\hat{\mathbf{P}}_N f_{\hat{\psi}_{\mathrm{CE}}} = 0$ μ-a.s., so it remains to show that $\hat{\psi}_{\mathrm{CE}} \to \psi_{\mathrm{CE}}$ in probability, which is implied by Theorem 3.4

> check

.

As

$$\|f_{\psi_1}(x) - f_{\psi_2}(x)\| = w(x)\|\dot{z}(\psi_1) - \dot{z}(\psi_2)\|$$

for all $\psi_1, \psi_2 \in \Psi$, $\mathbf{G}w < \infty$ and **(A2)** imply the local Lipschitz condition Equation (LL) in Theorem 3.3. Furthermore, by **(A3)** it holds

$$\mathbf{G}\|f_\psi\|^2 \leq \mathbf{G}w^2\|\dot{z}(\psi)\|^2 + 2\|\dot{z}(\psi)\|\mathbf{G}\|wT\| + \mathbf{G}\|wT\|^2 < \infty.$$

Additionally $\psi \mapsto \mathbf{G}f_\psi = (\mathbf{G}w)\dot{z}(\psi) + \mathbf{G}wT$ is differentiable everywhere, with Jacobian $(\mathbf{G}w)\ddot{z}(\psi)$, where $\ddot{z}(\psi) = \partial_\psi \dot{z}(\psi)$ is the Hessian of the cumulant generating function, which equals the negative Fisher information $-I(\psi_{\mathrm{CE}})$ as $\mathbf{G}_\psi, \psi \in \Psi$ form a natural exponential family and the regularity conditions **(A5)** allow differentiation under the integral. Thus we see that

$$\mathbf{G}f_{\psi_{\mathrm{CE}}} = \mathbf{P}\left(\dot{z}(\psi_{\mathrm{CE}}) + T\right) = \dot{z}(\psi_{\mathrm{CE}}) + \mathbf{P}T = 0,$$

by definition of ψ_{CE}, so

$$\mathrm{Cov}_{\mathbf{G}}\left(w(T - \nabla_{\psi_{\mathrm{CE}}}\log Z(\psi_{\mathrm{CE}}))\right) = \mathrm{Cov}_{\mathbf{G}}(f_{\psi_{\mathrm{CE}}}) = \mathbf{G}f_{\psi_{\mathrm{CE}}}f_{\psi_{\mathrm{CE}}}^T.$$

As $W = \frac{w}{\mathbf{G}w}$ By Equation (3.7) the asymptotic covariance matrix is

$$V_{\mathrm{CE}} = B_{\mathrm{CE}} M_{\mathrm{CE}} B_{ce}$$

which shows the asymptotic normality.

> consistency - additional conditions?

Estimating B_{CE} by $\hat{B}_{\mathrm{CE}} = I(\hat{\psi}_{\mathrm{CE}})$ and

$$M_{\mathrm{CE}} = \mathbf{G}W^2(T - \mathbf{G}_{\psi_{\mathrm{CE}}}T)(T - \mathbf{G}_{\psi_{\mathrm{CE}}}T)^T = \mathbf{P}W(T - \mathbf{G}_{\psi_{\mathrm{CE}}})(T - \mathbf{G}_{\psi_{\mathrm{CE}}})^T$$

by

$$\hat{M}_{\mathrm{CE}} = \hat{\mathbf{P}}_N W \left(T - \mathbf{G}_{\hat{\psi}_{\mathrm{CE}}}T\right)\left(T - \mathbf{G}_{\hat{\psi}_{\mathrm{CE}}}T\right)^T$$

yields the stated plug-in estimator. The promised consistency follows from **(A3)** and **(A4)**. □

Example 3.1 (univariate Gaussian). > ...

The form of the asymptotic covariance matrix is that of the sandwich estimator

> cite Huber

, corrected for the importance sampling with $\mathbf{G}$. This is not surprising: the CE-method performs maximum likelihood estimation where the data X_i

> sub or superskript

come from the misspecified $\mathbf{P}$. Additionally, we have to correct the variance for performing importance sampling with $\mathbf{G}$, instead of sampling directly from $\mathbf{P}$.

If $\mathbf{G}_\psi$ 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.

> expand a bit

The CE-method is routinely used for estimating failure probabilities for rare events [26]

> cite more

and has been applied to Bayesian inference [16, 17] and optimal control problems [31, 64].

3.4.3 Efficient importance sampling (EIS)

EIS[45] provides an alternative to the CE-method. Instead of minimizing the KL-divergence between the target $\mathbf{P}$ and $\mathbf{G}_\psi$, EIS aims at minimizing the variance of the logarithm of importance sampling weights.

> Chatterje paper

Thus, EIS finds ψ_{EIS} which solves

$$\psi_{EIS} = \mathrm{argmin}_{\psi \in \Psi} \mathrm{Var}_{\mathbf{P}}(\log w_\psi)$$
$$= \mathrm{argmin}_{\psi \in \Psi} \mathbf{P}(\log w_\psi - \mathbf{P}\log w_\psi)^2,$$

where $\log w_\psi = \log p - \log g_\psi$. As $\mathbf{P}\log w_\psi$ is usually intractable as well, we include it in the optimization problem, utilizing the fact that the mean is the minimizer of the squared distance functional. Here unnormalized weights $w \propto \frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{G}}$ may be used, as the unknown integration constant gets absorbed by the unknown mean. In total, EIS solves

$$(\psi_{\mathrm{EIS}}, \lambda_{\mathrm{EIS}}) = \mathrm{argmin}_{\psi \in \Psi, \lambda \in \mathbf{R}} \mathbf{P}(\log p - \log g_\psi - \lambda)^2.$$

Under the usual regularity conditions allowing to differentiate under the integral, the estimating equations for ψ_{EIS} read

$$\mathbf{P}\left(\left(\log p - \log g_\psi - \lambda\right)\nabla_\psi \log g_\psi\right) = 0 \tag{3.8}$$
$$\mathbf{P}\left(\log p - \log g_\psi - \lambda\right) = 0, \tag{3.9}$$

which we will use to derive asymptotics for $\hat{\psi}_{\text{EIS}}$.

Similar to the CE-method we restrict our in-depth analysis to natural exponential family proposals where $\log g_\psi(x) = \psi^T T(x) - \log Z(\psi) + \log h(x)$. In this case Equation (3.8) simplifies to

$$\mathbf{P}\left(\left(\log p - \psi^T T + \log Z(\psi) - \log h - \lambda\right)\left(T - \mathbf{G}_\psi T\right)\right) = 0,$$
$$\lambda = \mathbf{P}(\log p - \log g_\psi) = \mathcal{D}_{\text{KL}}\left(\mathbf{P}||\mathbf{G}_\psi\right).$$

As the first term is centered under $\mathbf{P}$, this is equivalent to $\log w_\psi$ and T being orthogonal in $L^2(\mathbf{P})$. Unfortunately, this formulation does not allow for an analytical solution of ψ_{EIS}, the problematic term being $\log Z(\psi)$, leading to an implicit equation for ψ. However, we can achieve an explicit equation by reparameterizing to $\lambda' = \lambda - \log Z(\psi)$, which results in a weighted linear least squares problem

$$\min_{\psi \in \Psi, \lambda' \in \mathbf{R}} \mathbf{P}\left(\log p - \log h - \psi^T T - \lambda'\right)^2.$$

Thus the optimal $(\psi_{\text{EIS}}, \lambda'_{\text{EIS}})$ are given by the best linear prediction of $\log p - \log h$ by the sufficient statistic T under $\mathbf{P}$. Therefore, if $\text{Cov}_\mathbf{P} T$ is non-singular,

do we need T in lambda?

$$\lambda'_{\text{EIS}} = \mathbf{P}\log p - \log h$$
$$\psi_{\text{EIS}} = \text{Cov}_\mathbf{P}\left(T\right)^{-1}\text{Cov}_\mathbf{P}\left(T, \log p - \log h\right). \tag{3.10}$$

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.

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

$$\left(\hat{\lambda}, \hat{\psi}_{\text{EIS}}\right) = \text{argmin}_{\lambda, \psi} \sum_{i=1}^N W^i\left(\log p(X^i) - \log g_\psi(X^i) - \lambda\right)^2,$$

where $X^1, \ldots, X^N \overset{\text{i.i.d}}{\sim} \mathbf{G}$. 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

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

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

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

and

$$W = \text{diag}\left(W_1, \ldots, W_N\right),$$

and the observations are

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

Alternatively, replacing $\mathbf{P}$ by $\hat{\mathbf{P}}_N$ in Equation (3.10), we obtain the equivalent formulation

$$\hat{\psi}_{\text{EIS}} = \text{Cov}_{\hat{\mathbf{P}}_N}(T)^{-1}\text{Cov}_{\hat{\mathbf{P}}_N}\left(T, \log p - \log h\right), \tag{3.11}$$

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 $\psi \in \Psi$ such that $\mathbf{P} = \mathbf{G}_\psi$, then EIS finds the optimal $\psi_{\text{EIS}} = \psi$ a.s. for a finite number of samples.

Proposition 3.2 (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 both Ψ and the support of the sufficient statistic* $\text{supp}\, T$ *are 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 $\log p - \log h = \psi_{\mathbf{P}}^T T - \log Z(\psi_{\mathbf{P}})$. Thus $\text{Cov}_{\hat{\mathbf{P}}_N}\left(T, \log p - \log h\right) = \text{Cov}_{\hat{\mathbf{P}}_N}(T)\,\psi_{\mathbf{P}}$. If we can show that $\text{Cov}_{\hat{\mathbf{P}}_N}T$ is non-singular, Equation (3.11) implies that $\hat{\psi}_{\text{EIS}} = \psi_{\mathbf{P}}$ a.s..

If $\text{Cov}_{\hat{\mathbf{P}}_N}T$ were singular, there would exist a $\psi \in \Psi$ such that $\text{Cov}_{\hat{\mathbf{P}}_N}\left(\psi^T T\right) = 0$, as Ψ is open and contains 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 $i = 1, \ldots, N$. As the weights are a.s. positive by the assumed equivalence of $\mathbf{G}$ and $\mathbf{P}$, the same holds true for $T(X^i), i = 1, \ldots, N$. If N is bigger than k, this is a contradiction to $\text{supp}\, T$ being open, so $\text{Cov}_{\hat{\mathbf{P}}_N}T$ is non-singular and the result is shown.

> think a bit more about this, what if **T** is not a homeo?

$\square$

> discuss applicability of CLT

Theorem 3.6 (asymptotic normality of $\hat{\psi}_{\text{EIS}}$)**.** *Let $\mathbf{G}_\psi$ form a natural exponential family with densities $g_\psi(x) = \frac{h(x)}{Z(\psi)}\exp\left(\psi^T T(x)\right)$ w.r.t. μ. Let $\mathbf{G}, \mathbf{P}$ be two other probability measures such that $\mathbf{G} \ll \mathbf{P}$ and let $W = \frac{d\mathbf{P}}{d\mathbf{G}}$ be the normalized importance sampling weights. Assume $\lambda(\psi) = \mathbf{P}\log w_\psi$ is known and the following conditions hold:*

> faster convergence if **P** is from same EF?

(B1) ...

(B2) $\mathcal{D}_{KL}\left(\mathbf{P}||\mathbf{G}_\psi\right) < \infty$ *for all ψ*

> prob. suffices locally

,

(B3) T *and* $\log w_{\psi_{EIS}}$ *are square integrable w.r.t.* $\mathbf{P}$*,*

(B4) *regularity conditions, and*

(B5) $Cov_{\mathbf{P}}(T)$ *is positive definite.*

WLOG, assume that $\mathbf{P}T = 0$. Then, as N goes to ∞,

$$\sqrt{N}\left(\begin{pmatrix}\hat{\lambda}\\\hat{\psi}_{EIS}\end{pmatrix} - \begin{pmatrix}\lambda\\\psi_{EIS}\end{pmatrix}\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, V_{EIS})$$

where

$$V_{EIS} = B_{EIS}M_{EIS}B_{EIS}$$

with

$$B_{EIS} = \begin{pmatrix} 1 & 0 \\ 0 & (Cov_{\mathbf{P}}T)^{-1} \end{pmatrix}$$

$$M_{EIS} = \mathbf{G}\left(W^2\left(\log\frac{p}{h} - \psi_{EIS}{}^T T - \lambda_{EIS}\right)^2 \begin{pmatrix} 1 & T^T \\ T & TT^T \end{pmatrix}\right).$$

In particular, the asymptotic variance of $\hat{\psi}_{EIS}$ is

$$(Cov_{\mathbf{P}}T)^{-1}\,\mathbf{G}\left(W^2\left(\log\frac{p}{h} - \psi_{EIS}{}^T T - \lambda_{EIS}\right)^2 TT^T\right)(Cov_{\mathbf{P}}T)^{-1}.$$

Proof. This proof follows the same strategy as that for Theorem 3.5 and uses the same notation (Notation 3.2). The estimating equations for λ and ψ_{EIS} are given by

$$x \mapsto f_{\lambda,\psi}(x) = \nabla_\lambda - \frac{1}{2}w(x)\left(\log\frac{p(x)}{h(x)} - \psi^T T(x) - \lambda\right)^2 = w(x)\left(\log\frac{p(x)}{h(x)} - \psi^T T(x) - \lambda\right)\begin{pmatrix} 1 \\ T(x) \end{pmatrix}.$$

For $\lambda_1, \lambda_2 \in \mathbf{R}$ and $\psi_1, \psi_2 \in \Psi$ the Lipschitz condition Equation (LL) are fulfilled, as

$$\|f_{\lambda_1,\psi_1} - f_{\lambda_2,\psi_2}\| = |w(x)|\,|(\lambda_2 - \lambda_1 + T(x)(\psi_2 - \psi_1))|\left\|\begin{pmatrix} 1 \\ T(x) \end{pmatrix}\right\|$$

$$\leq \underbrace{|w(x)|\left\|\begin{pmatrix} 1 & T^T(x) \\ T(x) & T(x)T^T(x) \end{pmatrix}\right\|}_{:=\dot{f}}\|(\lambda_2 - \lambda_1, \psi_2 - \psi_1)\|,$$

and $\mathbf{G}\dot{f} < \infty$ by (B3).

At the optimal $\lambda_{\text{EIS}}, \psi_{\text{EIS}}$ it holds

$$\mathbf{G}f_{\lambda_{\text{EIS}},\psi_{\text{EIS}}} = \mathbf{P}\left(\log\frac{p}{h} - \psi_{\text{EIS}}{}^T T - \lambda_{\text{EIS}}\right)\begin{pmatrix} 1 \\ T \end{pmatrix} < \infty,$$

as both T and $\log w(\psi_{\text{EIS}})$ are in $L^2(\mathbf{P})$ by (B3).

<div style="background:#E8871E;padding:2px 8px;border-radius:10px;display:inline-block">**L2 defnieren**</div>

By the assumed regularity conditions, $(\lambda, \psi) \mapsto \mathbf{G}f_{\lambda,\psi}$ is differentiable, with Jacobian

$$B_{\text{EIS}}^{-1} = \partial_{\lambda,\psi}\mathbf{G}f_{\lambda,\psi} = \mathbf{P}\begin{pmatrix} 1 \\ T \end{pmatrix}\begin{pmatrix} 1 & T^T \end{pmatrix}$$

$$= \mathbf{P}\left(\begin{pmatrix} 1 & T^T \\ T & TT^T \end{pmatrix}\right) = \begin{pmatrix} 1 & 0 \\ 0 & \text{Cov}_{\mathbf{P}}(T) \end{pmatrix},$$

as $\mathbf{P}T = 0$.

As $\hat{\psi}_{\text{EIS}}$ solves the estimating equations, we have $\hat{\mathbf{P}}_N f_{\hat{\psi}_{\text{EIS}}} = 0 = o_P\left(N^{-\frac{1}{2}}\right)$. It remains to show that $\hat{\psi}_{\text{EIS}} \to \psi_{\text{EIS}}$ in probability.

<div style="background:#E8871E;padding:2px 8px;border-radius:10px;display:inline-block">**show it**</div>

Finally, as $\mathbf{G}f_{\lambda_{\text{EIS}},\psi_{\text{EIS}}} = 0$,

$$M_{\text{EIS}} = \mathbf{G}\left(f_{\lambda_{\text{EIS}},\psi_{\text{EIS}}} f_{\lambda_{\text{EIS}},\psi_{\text{EIS}}}^T = \mathbf{G}w^2\left(\log\frac{p(x)}{h(x)} - \psi_{\text{EIS}}{}^T T - \lambda_{\text{EIS}}\right)^2 \begin{pmatrix} 1 & T^T \\ T & TT^T \end{pmatrix}\right)$$

$\square$

Theorem 3.7. [think about regularity conditions]

Under the appropriate regularity conditions, as N goes to ∞,

$$\sqrt{N}\left(\hat{\psi}'_{EIS} - \psi_{EIS}'\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, V_{EIS})$$

where

$$V_{EIS} = \left(\mathbf{P}(T'(T')^T)\right)^{-1}\left(\mathbf{G}\left(W^2(\log p - (T')^T\psi')^2 T'(T')^T)\right)\right)\left(\mathbf{P}(T'(T')^T)\right)^{-1}.$$

3.5 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 3.5.1) or approximating with a Gaussian discrete-time Markov process (Section 3.5.2).

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. At least for the models considered in Chapter 4, the targets arise as posterior distributions of LCSSM and as such they are unimodal and have non-heavy tails, i.e. there is a Gaussian distribution such that importance sampling is feasible

[rethink, is this really the case, for LCSSM it depends on log-partition function of observations]

. 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

[ref to chapter, check that I also present that there]

.

[small lit. review]

3.5.1 The GLSSM-approach

The first approach

[cite]

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

[cite]

algorithm. 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

$$
\begin{aligned}
\mathbf{G}_\psi &= \mathcal{L}(X|Z = z), \\
Z_t &= B_t X_t + \eta_t, \\
\eta_t &\sim \mathcal{N}(0, \Omega_t), \\
\Omega_t &= \operatorname{diag}(\omega_t^2) = \operatorname{diag}(\omega_{t,1}^2, \ldots, \omega_{1,p}^2).
\end{aligned}
\tag{3.12}
$$

where the distribution of X is given by (3.3), $\psi = (z, \omega^2)$ for $z = (z_0, \ldots, z_n) \in \mathbf{R}^{n \times m}$ and $\omega^2 = (\omega_0^2, \ldots, \omega_n^2) \in \mathbf{R}^{n \times m}$. Alternatively the natural parametrization $\psi = (z \oslash \omega^2, -1 \oslash (2\omega^2))$ 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, as

> add signal restriction above

$p(y_t|x_t)$ and $g(z_t|x_t)$ depend on x_t only through the signal $s_t = B_t x_t$, we have

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

which implies that auto-normalized weights may be calculated by using the signal smoother [29, Theorem 2].

> introduce LCSSM with linear signals above

As Durbin and Koopman [14, Section 4.5.3] argue, it is often computationally more efficient to treat only on the signals $(S_t)_{t=0,\ldots,n}$ instead of the states $(X_t)_{t=0,\ldots,n}$, 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

> add degernate case to gaussian conditional lemma

, with known conditional mean and covariance matrix and density $p(x|s) = g(x|s)$. If $(\tilde{X}_t)_{t=0,\ldots,n}$ 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) $\prod_{t=0}^n \frac{p(y_t|s_t)}{g(y_t|s_t)}$. 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 (3.12) 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 [14, 15] and may even be applied when the observation densities are not log-concave[29]. The approach also leads to efficient implementation for EIS [34]. 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

Algorithm 4 The LA for LCSSM

Algorithm 5 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 (3.12) 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 $\Omega = \text{block-diag}(\Omega_0, \ldots, \Omega_n)$. 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 (3.12) 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 $\tilde{\mu} = \mu + \Sigma(\Sigma + \Omega)^{-1}(z - \mu)$ and covariance matrix $\tilde{\Sigma} = \left(\Sigma^{-1} + \Omega^{-1}\right)^{-1}$ for $\Omega = \text{diag}(\omega^2)$, where $\omega^2 > 0$. Assuming that Σ is non-singular, we can reparameterize the objective function of the CE-method by $\tilde{\mu}$,

$$\max_{z, \omega^2} \int p(x|y) \log g_\psi(x|z) \mathrm{d}x = \max_{\tilde{\mu}, \omega^2} \int p(x|y) \left(-\frac{1}{2}(x - \tilde{\mu})^T\tilde{\Sigma}^{-1}(x - \tilde{\mu}) - \frac{1}{2}\log \det \tilde{\Sigma}\right) \mathrm{d}x$$
$$= \max_{\tilde{\mu}, \omega^2} -\frac{1}{2}(\gamma - \tilde{\mu})^T\tilde{\Sigma}^{-1}(\gamma - \tilde{\mu}) - \frac{1}{2}\text{trace}\left(\tilde{\Sigma}^{-1}\Gamma\right) - \frac{1}{2}\log \det \tilde{\Sigma},$$
(3.13)

where $\gamma = \mathbf{E}(X|Y = y)$ and $\Gamma = \text{Cov}(X|Y = y)$. Thus the optimal $\tilde{\mu}$ is γ and to find 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

$$\Gamma_{i,i} = \left(\left(\Sigma^{-1} + \text{diag}\left(\frac{1}{\omega_1}, \ldots, \frac{1}{\omega_p}\right)\right)^{-1}\right)_{i,i} = \left(\Sigma - \Sigma(\Sigma + \Omega)^{-1}\Sigma\right)_{i,i}$$
(3.14)

has to hold for all $i = 1, \ldots, p$, 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 (3.14) is not guaranteed to have a solution. For the i-th unit-vector $e_i \in \mathbf{R}^p$ we can reformulate Equation (3.14) to

$$\Sigma_{i,i} - \Gamma_{i,i} = e_i^T\Sigma^T(\Sigma + \Omega)^{-1}\Sigma e_i > 0$$
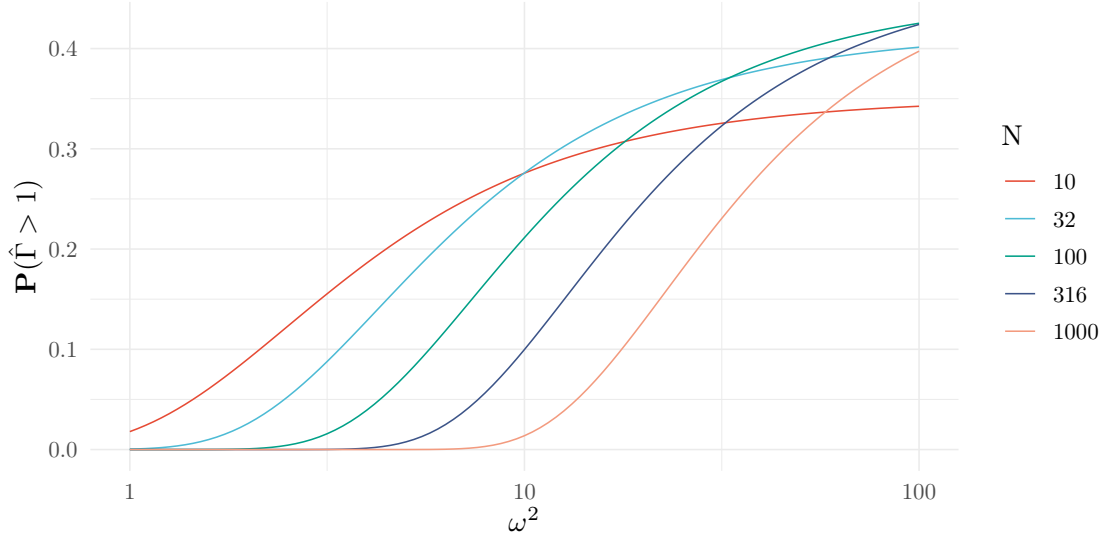
Figure 3.1: We show the probability that the estimated posterior variance $\hat{\Gamma}$ is bigger than the prior variance 1 when varying the noise variance ω^2. todo: ausführlicher beschreiben

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

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

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

> can we find example?

Second, even if there is an analytical solution Ω to Equation (3.14), 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 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 $X^1, \ldots X^N$ 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 (3.14) has no solution $\omega^2 \in \mathbf{R}_{\geq 0}$

> introduce symbol

. 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 $\chi^2_{N-1} \approx \mathcal{N}(N-1, 2(N-1))$ [28, Section 18.5]. We illustrate this in Figure 3.1, 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 (3.12) with practically relevant probability.

In higher-dimensional settings, e.g. when applying the CE-method to SSMs, we expect this phenomenon to occur even more often. In the extreme case of independent marginals, i.e. when Σ is a diagonal matrix, Equation (3.14) reduces to $(n+1)p$ many decoupled equations, where

$\hat{\Gamma}_{i,i}, i = 1, \ldots, (n+1)p$ 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 $((n+1)p, q)$ 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 (3.14) 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 (3.14) 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 our optimization, then $\Gamma > 1$ implies that the CE-method chooses this as the estimate, i.e. $\mathbf{G}_{\hat{\psi}_{\mathrm{CE}}}$ 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. [14, Section 4.10]. However, if there are a lot of failures, the optimally chosen $\mathbf{G}_{\hat{\psi}_{\mathrm{CE}}}$ will still 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 to apply the CE-method to SSMs.

3.5.2 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

$$
\begin{aligned}
\mathbf{G}_\psi &= \mathcal{L}(U + v), \\
v &\in \mathbf{R}^{(n+1)m}, \\
U_0 &\sim \mathcal{N}(0, R_0 R_0^T), \\
U_t &= C_t U_{t-1} + R_t \nu_t, \\
C_t &\in \mathbf{R}^{m \times m}, \\
\nu_t &\sim \mathcal{N}(0, I), \\
\mathbf{R}_t &\in \mathbf{R}^{m \times m} \text{lower triangular, positive diagonal}
\end{aligned}
\tag{3.15}
$$

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

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

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 in [45] 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, [34].

To perform importance sampling with $\mathbf{G}_\psi$ in model (3.15) we not only 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

$$
w(x) \propto \frac{p(y|x)p(x)}{g_\psi(x)} = \prod_{t=0}^n \frac{p(y_t|x_t)p(x_t|x_{t-1})}{g_\psi(x_t|x_{t-1})},
\tag{3.16}
$$

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 (3.15) 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 Lemmas show. For a general treatment, we refer the reader to [36, Chapters 3 and 5].

Lemma 3.3. *Let* (X, Y) *be jointly Gaussian with distribution* $\mathcal{N}(\mu, \Sigma)$ *where*

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

and

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

for non-singular Σ. *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), *then* $Cov(X|Y) = P_{XX}^{-1}$.

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

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

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}. $\square$

Lemma 3.4. *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

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

Now $X \perp Y|Z$ is equivalent to $\text{Cov}((X, Y)|Z)$ 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 3.3, the sub-matrix of P whose rows and columns correspond to X and Y. $\square$

Applying Lemma 3.4 to model (3.15), 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 $t = 1, \ldots, n$.

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 from the theory of Gaussian-Markov-Random-fields (GMRF), i.e. Gaussian models whose dependency structure is given by a graph, which is here slightly adapted to our setting and notation.

Theorem 3.8 ([22, 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* $0 \le t < s \le n$ *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 3.8 we immediately obtain the following:

Corollary 3.1 (sparsity of L in model (3.15))**.** *Let $U \sim \mathbf{G}_\psi$ as in Equation (3.15), $P \succ 0$ be the precision matrix of $\overleftarrow{U} = (U_n, \ldots, U_0)$ 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:*

$$
L = \begin{pmatrix}
L_{n,n} & 0 & \cdots & \cdots & \cdots & 0 & 0 \\
L_{n-1,n} & L_{n-1,n-1} & 0 & \cdots & \cdots & 0 & 0 \\
0 & L_{n-2,n-1} & L_{n-2,n-2} & 0 & \cdots & 0 & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots & 0 & 0 \\
0 & 0 & 0 & \cdots & L_{1,2} & L_{1,1} & 0 \\
0 & 0 & 0 & \cdots & 0 & L_{0,1} & L_{0,0}
\end{pmatrix}, \tag{3.17}
$$

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

From L in Corollary 3.1 we obtain an iterative method of sampling from $\mathbf{G}_\psi$: If $v + U \sim \mathbf{G}_\psi$, then, as $\text{Cov} U = \left(LL^T\right)^{-1} = L^{-T}L^{-1}$, 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 $\overleftarrow{Z} = (Z_n, \ldots, Z_0) \sim \mathcal{N}(0, I)$. 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 $t = 1, \ldots, n$

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

Rearranging terms, we end up with the a Markov-process

$$
U_t = L_{t,t}^{-T} L_{t-1,t}^T U_{t-1} + L_{t,t}^{-T} Z_t, \tag{3.18}
$$

where Z_t is, by construction, independent of U_t. Thus for model (3.15), we obtain

$$
\begin{aligned}
R_t &= L_{t,t}^{-T} \text{ for } t = 0, \ldots, n, \\
C_t &= L_{t,t}^{-T} L_{t-1,t}^T \text{ for } t = 1, \ldots, n.
\end{aligned} \tag{3.19}
$$

> **improve: order to $\mathcal{O}(m^3)$**
>
> show that if covariance is SPD, then there is a unique MP with those marginal covariances (just calculate C and R). Then this MP is the optimal one, much faster to calculate (Cholesky decomposition of cov(xt, xt-1))

Here we see why we chose to use $\overleftarrow{U}$ in Corollary 3.1: had we applied Theorem 3.8 to U directly yields a Markov process in reverse time.

We now turn our attention to applying the CE-method to model (3.15). Following a similar argument as in the discussion surrounding Equation (3.13), we see that it suffices to choose P, the precision matrix of U, such that it minimizes

$$
\frac{1}{2} \text{trace}\left(P\hat{\Gamma}\right) - \frac{1}{2} \log \det P \tag{3.20}
$$

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

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

Here P is restricted to precision matrices that may arise in model (3.15), i.e., by Corollary 3.1, $P = LL^T$ where L possess structure as in (3.17). At first glance, this problem seems more involved than solving Equation (3.14): 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 (3.20). 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 $\text{Cov}(X_t, X_{t-1})$ and $\text{Cov}(X_0)$ 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 3.9 ([53, Theorem 2.1]). *Let Γ be a positive-definite matrix of size $n \times n$. Given a lower-triangular sparsity set $S \subset \{1, \ldots, n\}^2$, i.e. $i \geq j$ for all $(i,j) \in S$, let*

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

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

$$L_{s_i,i} = \frac{\Gamma_{s_i,s_i}^{-1} e_1}{\sqrt{e_1^T \Gamma_{s_i,s_i}^{-1} e_1}}, \tag{3.21}$$

where $s_i = \{j : (i,j) \in S\}$, 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.

For the problem at hand, the sparsity pattern $\mathcal{S}$ is given by the non-zero entries of L depicted in Equation (3.17) and the matrices Γ_{s_i,s_i} are sub-matrices of the covariances $\text{Cov}\left((X_{t-1}, X_t)\,|Y\right)$, for $t = 1, \ldots, n$ and $\text{Cov}(X_0|Y)$, i.e. we let $\Gamma = \text{Cov}\left(\overset{\leftarrow}{X}|Y\right)$ in Theorem 3.9.

Let $l_t^j \in \mathbf{R}^{2m \times m}$ be the j-th column of

$$\begin{pmatrix} L_{t,t} \\ L_{t-1,t} \end{pmatrix},$$

where $j \in \{1, \ldots, m\}$. As $L_{t,t}$ is lower triangular, the first $j-1$ entries of l_t^j are 0. We obtain the remaining nonzero entries by computing

$$\frac{\Gamma_{t,j}^{-1} e_j}{\sqrt{e_j^T \Gamma_{t,j}^{-1} e_j}},$$

for $\Gamma_{t,j} \in \mathbf{R}^{(2m-(j-1)) \times (2m-(j-1))}$ the joint covariance matrix of the last $m - (j-1)$ components of X_t and all entries of X_{t-1}, conditional on $Y = y$, and $e_j \in \mathbf{R}^{2m-(j-1)}$ the first unit vector.

Putting everything together, we may apply the CE-method to estimate ψ in model (3.15) in the following way: Given importance samples $U^1, \ldots, U^N$ for $\mathcal{L}(X|Y = y)$ and associated unnormalized weights $w^1, \ldots, w^N$, we estimate v by

$$\hat{v} = \frac{\sum_{i=1}^N w^i X^i}{\sum_{i=1}^N w^i} \tag{3.22}$$

and the empirical covariance matrices

$$\widehat{\text{Cov}}\,(X_t, X_{t-1}) = \frac{\sum_{i=1}^N w^i (X_{t:t-1}^i - \hat{v}_{t-1:t})(X_{t:t-1}^i - \hat{v}_{t-1:t})^T}{\sum_{i=1}^N w^i} \tag{3.23}$$

$$\widehat{\text{Cov}}\,(X_0) = \frac{\sum_{i=1}^N w^i (X_0^i - \hat{v}_0)(X_0^i - \hat{v}_0)^T}{\sum_{i=1}^N w^i}. \tag{3.24}$$

We then determine $\hat{L}$ from these covariance matrices using Theorem 3.9. These steps are summarized in Algorithm 6.

Algorithm 6 The CE-method for the Markov proposal (3.15)

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 $U^1 + v^l, \ldots, U^N + v^l \overset{\text{i.i.d}}{\sim} \mathbf{G}_{\hat{\psi}^l}$ with fixed seed   ▷ Equation (3.15)
4:     determine unnormalized weights $w^1, \ldots, w^N$ ▷ Equation (3.16)
5:     estimate $\hat{v}^{l+1}$ ▷ Equation (3.22)
6:     estimate $\widehat{\text{Cov}}(U_t, U_{t-1}), t = 1, \ldots, n, \widehat{\text{Cov}}(U_0)$ ▷ Equation (3.23)
7:     determine $\hat{L}^{l+1}$ ▷ Theorem 3.9
8:     determine C^{l+1} and R^{l+1} from $\hat{L}^{l+1}$ ▷ Equation (3.19)
9:     set $\hat{\psi}^{l+1} = \left(\hat{v}^{l+1}, C^{l+1}, R^{l+1}\right)$
10:    set $l = l + 1$
11: **until** $\hat{\psi}^l$ converged
12: **return** $\hat{\psi}_{\text{CE}} = \hat{\psi}^l$

To run Algorithm 6 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 $X^1, \ldots, X^N$ from the LA and performing steps 5 to 9 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 $\text{Cov}\,(X_t, X_{t-1}|Z = z)$ and the marginal covariance $\text{Cov}\,(X_0|Z = z)$ can be computed as well.

The convergence criteria in Algorithm 6 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.

why not on cholesky root of Ξ?

In Line 3 we use the standard praxis of common random numbers (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 6 in Table 3.1 The total time complexity of single iteration of Algorithm 6 is $\mathcal{O}\left(N\,n\,m^2 + n\,m^4\right)$ 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 (3.15), where each simulation requires $\mathcal{O}(n)$ matrix-vector multiplications of dimension m.

Line 4 To evaluate the weights, Equation (3.16), 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, making, so this step has time-complexity $\mathcal{O}(m^2)$. In Equation (3.16) 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.

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 L (Line 7)	$\mathcal{O}(n\,m^4)$	$\mathcal{O}(n\,m^2)$
determining C and R (Line 8)	$\mathcal{O}(n\,m^3)$	$\mathcal{O}(n\,m^2)$

Table 3.1: Time and space complexities of individual steps in Algorithm 6.

Line 5 Calculating the weighted mean $\hat{v} \in \mathbf{R}^{(n+1)m}$, Equation (3.22), requires $\mathcal{O}(N\,n\,m)$ operations.

Line 6 Calculating the weighted covariance matrices, Equation (3.23), requires $(n+1)$ times N many $m \times 1$ with $1 \times m$ vectors.

Line 7 To determine $L_{t,t}$ and $L_{t-1,t}$ we have to solve m times the linear systems of equations given in Equation (3.21), where the dimension of the system is $2m, \ldots, m+1$. This requires $\mathcal{O}(m^4)$ many operations, and we have to perform it for every one of the $n+1$ time points. The result is L with sparsity structure given by Equation (3.17), which has $\mathcal{O}(nm^2)$ many non-zero entries.

Line 8 For each of the $\mathcal{O}(n)$ many C_t and R_t we have to invert a triangular matrix of dimension m.

An efficient implementation of Algorithm 6 can improve on some of the other steps involved. There is no need to calculate the C_t and R_t matrices explicitly, instead we can calculate C_tU_{t-1} efficiently by solving the linear system $L_{t,t}^T (C_tU_{t-1}) = L_{t-1,t}^T U_{t-1}t$ by back-substitution, as $L_{t,t}^T$ is an upper triangular matrix. Similarly, to compute R_tZ_t we solve $L_{t,t}^T (R_tZ_t) = Z_t$ by back-substitution. However, the main bottleneck for the time-complexity of Algorithm 6 is determining L with time complexity $\mathcal{O}(n\,m^4)$, which we have not been able to improve upon.

gains by succesive block matrix inversion?

If the main bottleneck for space lies in the $\mathcal{O}(N\,n\,m)$ simulation part, we may reduce this by simulating twice from model (3.15) 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 from time t and $t+1$, i.e. $\mathcal{O}(N\,m)$ space. This reduces the space complexity to $\mathcal{O}(N\,m + n\,m^2)$.

We demonstrate these improvements in Algorithm 7. Additionally, we calculate the weights on the log scale for numerical stability.

The advantage of Algorithms 6 and 7 over applying the CE-method to the GLSSM model (3.12) 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 3.5.1. Additionally, determining the optimal parameters (v, C, R) or (v, L) is numerically stable, involving only inversion of small matrices. Compare this with solving Equation (3.14), 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 (3.15), generating N samples requires only $\mathcal{O}(N\,n\,m^2)$ operations, whereas sampling from model (3.12) 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 at least as many samples as there are dimensions, 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 L becomes more time-intensive. Additionally, the dimension of the parameter ψ

Algorithm 7 Time and space improved version of Algorithm 6. Instructions involving the free index i are to be performed for all $i = 1, \ldots, N$ samples.

Require: LCSSM (Definition 3.5), observations Y, initial estimate $\hat{\psi}^0 = \left(v^0, L^0\right)$, sample size N

 set $l = 0$

 repeat

  simulate $Z_0^1, \ldots, Z_0^N \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, I)$

  set $U_0^i = (L_{0,0}^l)^{-T} Z_0^i$ $\triangleright$ backsubstitution

  set $X_0^i = v_0^l + U_0^i$

  set $\log w^i = \log p(y_0|X_0^i) + \log p(X_0^i) + \frac{1}{2}\|Z_0^i\|^2$ $\triangleright \log g(X_0^i) = -\frac{1}{2}\|Z_0^i\|_2^2 + C$

  store current RNG state

  for $t \leftarrow 1, \ldots, n$ **do**

   simulate $Z_t^1, \ldots, Z_t^N \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, I)$

   set $U_t^i = (L_{t,t}^l)^{-T}(L_{t-1,t}^l)^T U_{t-1}^i + (L_{t,t}^l)^{-T} Z_t^i$ $\triangleright$ backsubstitution

   set $X_t^i = v_t^l + U_t^i$

   set $\log w^i = \log w^i + \log p(y_t|X_t^i) + \log p(X_t^i|X_{t-1}^i) + \frac{1}{2}\|Z_t^i\|^2$

  end for

  set $\log w^i = \log w^i - \max_{i=1,\ldots,N} \log w^i$ $\triangleright$ ensure $\log w^i \leq 0$

  set $w^i = \exp(\log w^i)$

  set $W^i = \frac{w^i}{\sum_{i=1}^N w^i}$ $\triangleright$ auto-normalized weights

  set $v_0^{l+1} = \sum_{i=1}^N W^i X_0^i$

  restore RNG state

  for $t \leftarrow 1, \ldots, n$ **do**

   simulate $Z_t^1, \ldots, Z_t^N \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, I)$

   set $U_t^i = (L_{t,t}^l)^{-T}(L_{t-1,t}^l)^T U_{t-1}^i + (L_{t,t}^l)^{-T} Z_t^i$ $\triangleright$ backsubstitution

   set $X_t^i = v_t^l + U_t^i$

   set $v_t^{l+1} = \sum_{i=1}^N W^i X_t^i$

   set $\widehat{\text{Cov}}(X_{t-1}, X_t) = \sum_{i=1}^N W^i \left(X_{t-1:t}^i - v_{t-1:t}^l\right)\left(X_{t-1:t}^i - v_{t-1:t}^l\right)^T$

  end for

  determine $\hat{L}^{l+1}$ $\triangleright$ Theorem 3.9

  set $\hat{\psi}^{l+1} = \left(\hat{v}^{l+1}, \hat{L}^{l+1}\right)$

  set $l = l + 1$

 until $\hat{\psi}^l$ converged

 return $\hat{\psi}_{\text{CE}} = \hat{\psi}^l$

---

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.

To improve the speed of computation of both algorithms, L has to be computed faster.

> speculate how to do this, can we directly read the MP off-of covariances? need proof!

3.6   Accouting for multimodality and heavy tails

Performing importance sampling with the Gaussian models discussed so far will work well only if the smoothing distribution $p(x|y)$ is well approximated by a Gaussian distribution. However, a Gaussian distribution is a very specific kind of distribution, in particular, it is an unimodal distribution and has light tails

> check for correct wording

.

If the smoothing distribution violates any of these assumptions, importance sampling with the models presented so far is likely to fail, i.e. requiring large sample sizes for both finding the optimal importance sampling parameter $\hat{\psi}$ as well as the final importance sampling evaluation.

There are however techniques to keep most of the computational efficiency discussed in the above sections to address both multimodality as well as heavy tails.

We start with heavier than gaussian tails: the textbook example of a heavy tailed distribution is the multivariate t-distribution with density

$$\ldots.$$

for degrees of freedom $\nu > 1$

> ?

, location μ and scale matrix Σ. When $\nu > 2$ then this distribution has mean μ and if $\nu > 3$ it has covariance matrix ?

> check

.

The main properties necessary to facilitate Gaussian importance sampling strategies above are that the distribution $p(x|y)$ is analytically tractable and simulation from it is possible. These properties still hold for the multivariate t-distribution and, in fact, for the even larger class of elliptical distributions:

Theorem 3.10 (Conditional distribution of elliptical distributions). > cite the correct book

As one can readily see from Theorem 3.10 the parameters of the smoothing distribution $p(x|y)$ if $p(x, y)$ follows an elliptical distribution is again elliptical and its parameters only depend on quantities that are computed by the Kalman smoother.

> elaborate

> present some models with heavy tails

3.7   Maximum likelihood estimation in SSMs

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 5 and 7 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.

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 3.2 (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 $x^* = 0$. The LA is $\mathbf{G}_{\mathrm{LA}} = \mathcal{N}\left(0, \frac{1}{\varepsilon^2 - \varepsilon + 1}\right)$, 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 fails.

The CE-method minimizes the KL-divergence between $\mathbf{P}$ and $\mathbf{G}_\psi$, is given by $\mathbf{G}_{\mathrm{CE}} = \mathcal{N}(0,\sigma^2)$, 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 $\mathbf{G}_{\mathrm{CE}}$ is consistent.

add proof for $\frac{1}{2}$ to appendix

EIS does not yield analytically tractable proposals in this setting. We will investigate its performance for this scale mixture in Example 3.4.

do so

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 [45, 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 [**Branch1999Subspace**] or the BVLS (bounded variable least squares) solver [**Stark1995Boundedvariable**] 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 3.5.1 for the CE-method.

The CE-method presented in Section 3.5.2 (Algorithm 7) depends on the fact that the covariance matrix of the posterior $\text{Cov}(X|Y = y)$ 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 7 provides a globally optimal solution, though it may not be unique.

3.8.2 Time and space complexity

As the LA is based on the Kalman-smoother, the time-complexity of a single iteration is $\mathcal{O}(n\,m^3)$, which may be improved to $\mathcal{O}(n\,p^3)$ if the efficient signal smoother is applicable. 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 3.5.2). For EIS, using the forwards filter, backwards sampling (FFBS) requires $\mathcal{O}(N\,n\,m^2 + n\,m^3)$ operations: the innovations share the same covariance matrices, so their Cholesky roots may be computed for all samples at once. While simulation is more expensive for EIS, obtaining the mode parameters is more 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. In total

3.8.3 Asymptotic variance

Example 3.3 (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 3.5 and 3.6).

By symmetry, both ψ_{CE} and ψ_{EIS} are equal to 0. Then $I(\psi) = 1$ for all ψ, so

$$V_{\text{CE}} = \text{Cov}_{\mathbf{P}}(T) = \frac{\tau^2}{\sigma^2}, \tag{3.25}$$

where $\tau^2 = \mathbf{P}\,\text{id}^2$ is the second moment of $\mathbf{P}$. Additionally, $B_{\text{EIS}} = (\text{Cov}_{\mathbf{P}}(T))^{-1} = \frac{\sigma^2}{\tau^2}$ and

$$
\begin{aligned}
M_{\text{EIS}} &= \text{Cov}_{\mathbf{P}}\left(\left(\log\frac{p(x)}{h(x)} - \lambda_{\text{EIS}}\right)T\right) \\
&= \text{Cov}_{\mathbf{P}}\left(\left(\log p - \log h - \mathbf{P}(\log p - \log h)\right)T\right) \\
&= \frac{1}{\sigma^2}\int p(x)x^2\left(\log p(x) + \frac{x^2}{2\sigma^2} - \mathbf{P}\left(\log p(x) + \frac{\tau^2}{2\sigma^2}\right)\right)^2 \mathrm{d}x.
\end{aligned}
$$

Thus

$$V_{\text{EIS}} = B_{\text{EIS}}M_{\text{EIS}}B_{\text{EIS}} = \sigma^2\frac{\gamma}{\tau^4},$$

where $\gamma = \int p(x)x^2\left(\log p(x) + \frac{x^2}{2\sigma^2} - \mathbf{P}(\log p(x) + \frac{\tau^2}{2\sigma^2})\right)^2 \mathrm{d}x$.

Normal distribution   If $\mathbf{P} = \mathcal{N}(0, \tau^2)$ is a normal distribution, this reduces to

$$V_{\text{EIS}} = \frac{5}{2}\left(\frac{\tau^2}{\sigma^2} - 1\right)^2\frac{\sigma^2}{\tau^2} = \frac{5}{2}\frac{(V_{\text{CE}} - 1)^2}{V_{\text{CE}}}$$
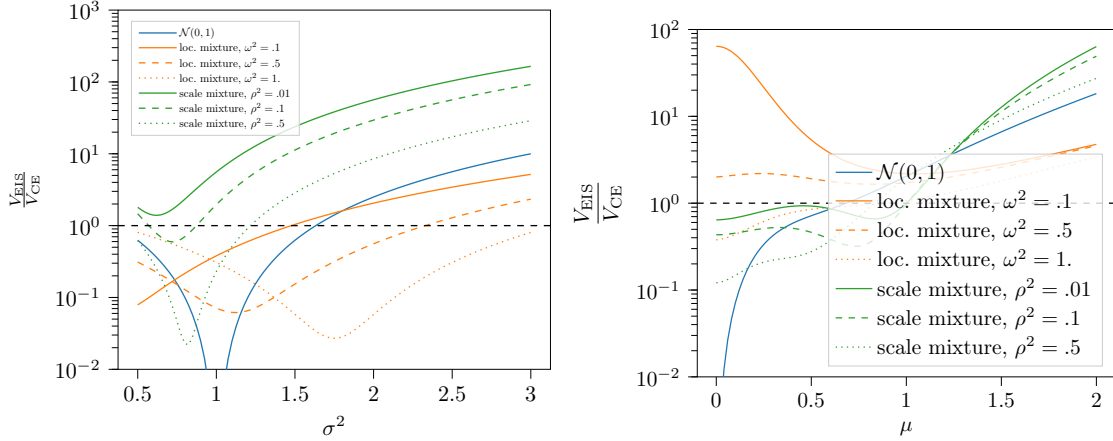
Figure 3.2: Asymptotic relative efficiency $\frac{V_{\mathrm{EIS}}}{V_{\mathrm{CE}}}$ for the normal distribution from Example 3.3 (left hand side) and Example 3.4 (right hand side). Here $\mathbf{P} = \mathcal{N}(0,1)$ is the standard normal distribution and $\mathbf{G} = \mathcal{N}(\mu, \sigma^2)$, 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 todo: reference to example where EIS is exact.. As

and so for $\tau^2 = \sigma^2$ $\hat{\psi}_{\mathrm{EIS}}$ converges faster than the standard $\mathcal{O}(N^{-\frac{1}{2}})$ rate. Indeed in this case $\hat{\psi}_{\mathrm{EIS}} = \psi_{\mathrm{EIS}}$ a.s. for $N > 1$ see

> write example for EIS being exact for exponential families

.

For importance sampling to be consistent, it is necessary that $\sigma^2 > \frac{\tau^2}{2}$. The left-hand side of Example 3.3 displays the behavior of the relative efficiency

$$\frac{V_{\mathrm{EIS}}}{V_{\mathrm{CE}}} = \frac{5}{2} \frac{(V_{\mathrm{CE}} - 1)^2}{V_{\mathrm{CE}}^2} = \frac{5}{2} \left(1 - \frac{2}{Vsdf_{\mathrm{CE}}} + \frac{1}{V_{\mathrm{CE}}^2} \right)$$

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 $\tau^2 = 1 + \omega^2 = -\frac{1}{2\psi_{\mathrm{CE}}}$. 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.

> interpretation: if variance set too low, CE better. if variance too big EIS better

> add scale mixture similar to example LA failure before

Example 3.4 (univariate Gaussian, μ fixed)**.** Consider the same setup as in Example 3.3, 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 $\mathrm{Cov}_{\mathbf{P}}T = \nu - \tau^4 + 4\tau^2\mu^2$ where $\nu = \mathbf{P}\,\mathrm{id}^4$ and $\tau^2 = \mathbf{P}\,\mathrm{id}^2$.

By matching moments, we obtain $\psi_{\mathrm{CE}} = -\frac{1}{2(\tau^2 + \mu^2)}$ and $I(\psi_{\mathrm{CE}}) = \frac{1}{2\psi_{\mathrm{CE}}^2} = 2(\tau^2 + \mu^2)^2$. In total

$$V_{\mathrm{CE}} = \frac{1}{4(\tau^2 + \mu^2)^4}\left(\nu - \tau^4 + 4\tau^2\mu^2\right) \tag{3.26}$$

For EIS,

$$
\begin{aligned}
\psi_{\text{EIS}} &= (\text{Cov}_{\mathbf{P}} T)^{-1} \text{Cov}_{\mathbf{P}} (T, \log p) \\
&= \left(\nu - \tau^4 + 4\tau^2\mu^2\right)^{-1} \underbrace{\int p(x)((x-\mu)^2 - \tau^2 - \mu^2)(\log p(x) - \mathbf{P}\log p(x))\, \mathrm{d}x}_{=\gamma}.
\end{aligned}
$$

Then

$$
V_{\text{EIS}} = \left(\nu - \tau^4 + 4\tau^2\mu^2\right)^{-2} \mathbf{P}\left((\text{id}-\mu)^4 \left(\log p - \psi_{\text{EIS}}(\text{id}-\mu)^2 - \mathbf{P}\log p + \psi(\tau^2 + \mu^2)\right)^2\right).
$$

Normal distribution   Consider now 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 accompanying source code for their calculation in sympy

ref it

.

3.8.4   Numerical convergence

3.8.5   ESS at the optimum

stress that cem gives global optimum, eis only approximate

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 3.3 and 3.4. 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 3.5 and 3.6 is not possible. 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 3.5 (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 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 model is given by

$$
\begin{aligned}
X_0 &\sim \mathcal{N}(0, \Sigma) \\
X_t &= \alpha X_{t-1} + \varepsilon_t \\
\varepsilon_t &\overset{\text{i.i.d}}{\sim} \mathcal{N}(0, (1 - \alpha^2)\Sigma), t = 1, \ldots, n
\end{aligned}
$$

where the $\varepsilon_1, \ldots, n$ and X_0 are jointly independent. The observations follow a conditional negative binomial distribution

$$
Y_t^i | X_t \sim \text{NegBinom}\left(\exp X_t^i, r\right), i = 1, \ldots, p
$$

where the parametrization is the one by mean and overdispersion parameter $r > 0$

> ref it

and individual observations are conditionally independent given the current state.

Our first simulation study concerns the non-asymptotic behavior of the CE-method and EIS estimators, i.e. finite sample analogs of Theorems 3.5 and 3.6. To this end, we let $m = 1$ in Example 3.5 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 $\hat{\mathbf{P}}_{\text{CE}}^N$ and $\hat{\mathbf{P}}_{\text{EIS}}^N$. 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 mean and Cholesky root of the precision matrix(CE-method). This procedure is repeated M times for every sample size N considered, with different initial random seeds, obtaining $\hat{\mathbf{P}}_{\text{CE}}^{N,i}$ and $\hat{\mathbf{P}}_{\text{EIS}}^{N,i}$ for $i = 1, \ldots, M$.

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 $\mu_i, \sigma_i^2 \in \mathbf{R}^{(n+1)\cdot m}$ be the marginal means and variances of $\mathbf{G}_{\text{CE}}^{N,i}$ ($\mathbf{G}_{\text{EIS}}^{N,i}$) for $i = 1, \ldots, M$. Now

$$
\widehat{\text{ASE}}_i = \frac{1}{(n+1)m}\left(\|\mu - \mu_i\|^2 + \|\sigma^2 - \sigma_i^2\|^2\right)
$$

is an estimate of the state-average squared error and

$$
\widehat{\text{AMSE}} = \frac{1}{M}\sum_{i=1}^{M}\widehat{\text{ASE}}_i
$$

is an estimate of the state-average mean squared error. The ASE_i is of interest to the practitioner as they usually only run a single iteration of the optimal importance sampling procedure. So while a low AMSE is desirable, the variance of ASE should also be small in practice, as otherwise several runs of the optimal importance sampling procedure may be required to obtain a good proposal.

In Figure 3.3 we show the $\widehat{\text{ASE}}_i$ for $i = 1, \ldots, M$ for both the CE-method and EIS. As is evident from this Figure, the CE-method consistently has a larger average mean squared error than EIS, for all values of N. Thus the CE-method requires several orders of magnitude more samples to obtain the same error as EIS.
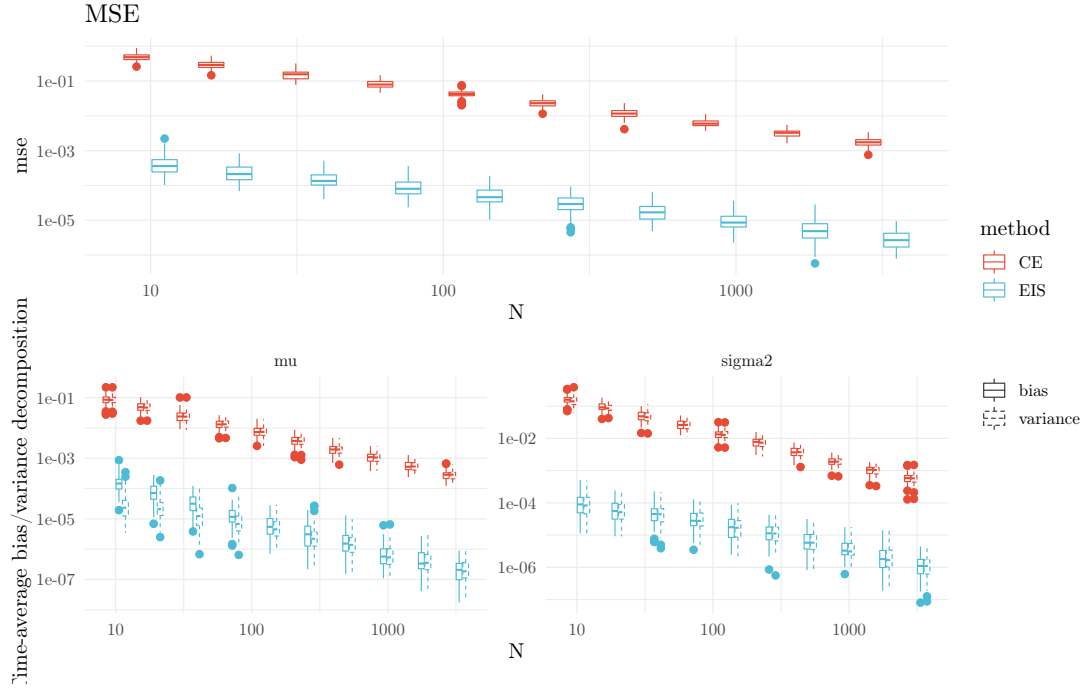
Figure 3.3: **TODO: ticks at the Ns**

`more interpretation`

For further investigation we perform a bias-variance decomposition of the A(M)SE for both the means μ and variances σ^2. Consider the averages means and variances over the M simulations,

$$\bar{\mu} = \frac{1}{M} \sum_{i=1}^{M} \mu_i \qquad\qquad \bar{\sigma}^2 = \frac{1}{M} \sum_{i=1}^{M} \sigma_i^2,$$

and the state-average squared bias and variance

$$\text{aBias} = \frac{1}{(n+1)m} \|\mu - \bar{\mu}\|^2$$

$$\text{aVar} = \frac{1}{M-1} \frac{1}{(n+1)m} \sum_{i=1}^{M} \|\bar{\mu} - \mu_i\|^2.$$

For the performance of importance sampling the efficiency factor $\text{EF} = \frac{\text{ESS}}{N}$ plays an important role, see

`ref correct chapter`

. 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} for the second moment of importance sampling weights ρ. Consequently, we also estimate EF for each of the M runs, using the same number of samples N as was used to determine the optimal parameter. We display the resulting efficiency factors in Figure 3.4. 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.
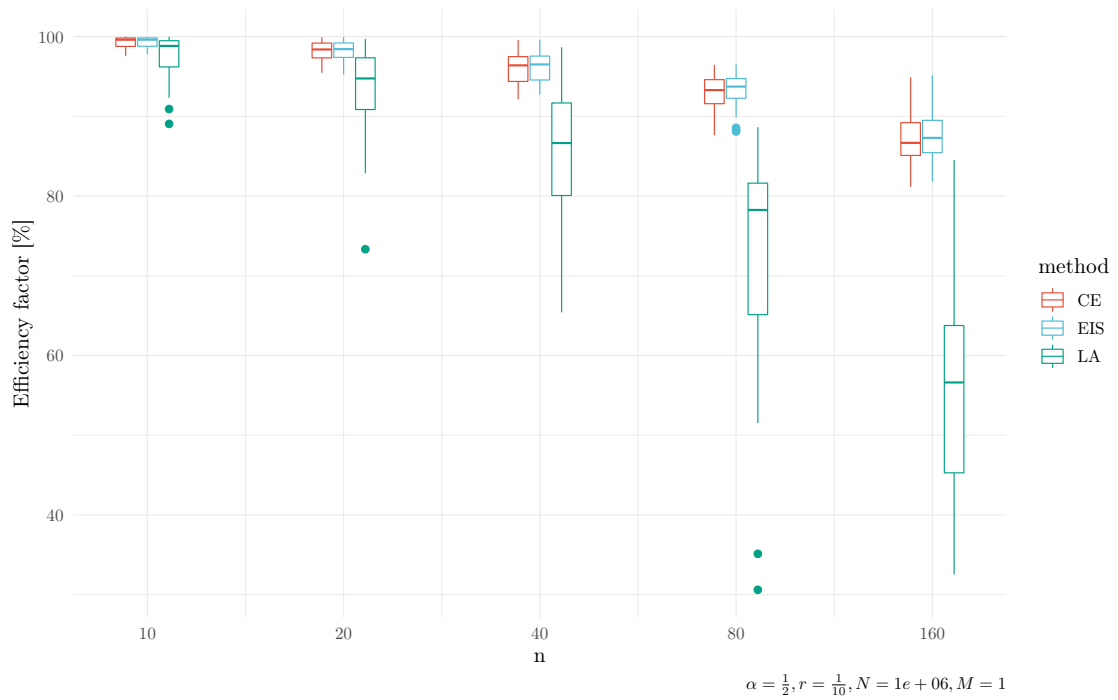
Figure 3.4: The efficiency factor degenerates as the number of time steps n increases. We show the estimated efficiency factor over 100 replications of estimating the optimal parameters for Example 3.5 with the CE-method and EIS with $N_{\text{true}} = 10^6$ and the resulting estimated efficiency factors 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

Chapter 4

Analysis of selected models

Contributions of this chapter

- ...

4.1   Spatial reproduction number model

1. essentially the Regional model presented in ECMI

4.2   Regional growth factor model

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 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 authority, the Robert Koch-Institut (RKI), via Github [46]. 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 [30], delays in reporting for AIDS [37, 63] and other infectious diseases [19]. Popular statistical approaches include methods from survival analysis [37] and generalized linear regression [63]. 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.
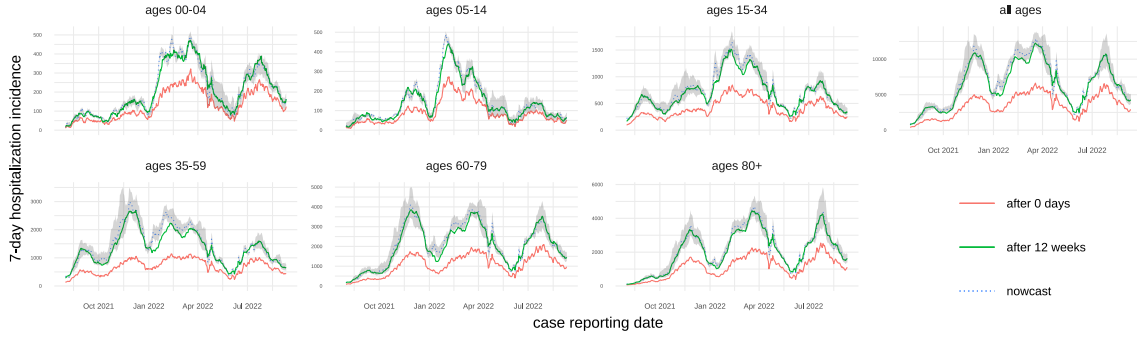
Figure 4.1: Germany's 7-day hospitalisation incidence changes due to various delays such as time to hospitalisation and delays in reporting. This figure shows the extent of these delays: incidences reported at the present date (red lines) severely underestimate the hospitalisation incidence (green solid 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.
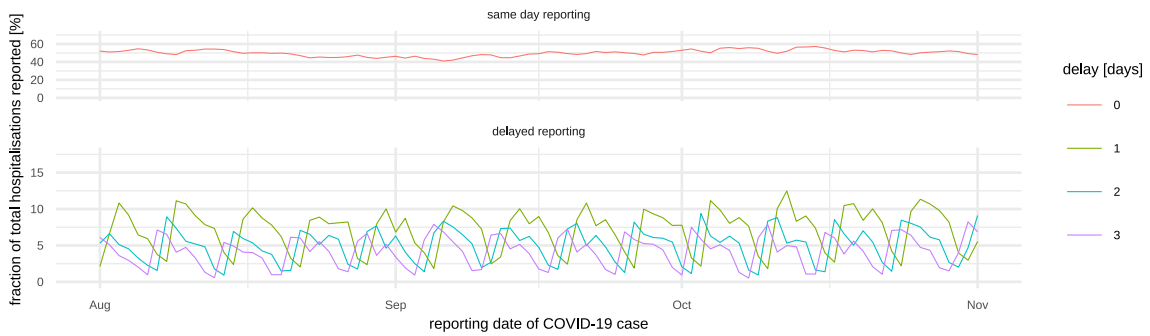
Figure 4.2: The hospitalisation incidence contains a double weekday effect owed to the reporting of both the COVID-19 case and the subsequent hospitalisation. While the weekday effect of the case reporting date is somewhat mitigated by summing over 7 day periods, the weekday effect of reporting date 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.

This has been studied with frequentist [39] and Bayesian [3, 25] methods. The generalized linear regression approach has origins in the chain ladder model from actuarial sciences [44] 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 of real-time nowcasting, it is often beneficial to incorporate epidemic dynamics into the model, this can be achieved by splines [25, 60] or by a latent process of infections [38].

Nowcasting methods have wide application in accouting for reporting delays [39], early outbreak detection [5, 52], and, in the recent COVID-19 epidemic, improving real-time monitoring of epidemic outbreaks [2, 3, 23, 54]. Evaluating a forecasting model in a real-time public health setting is advantageous as it avoid hindsight bias [13], however nowcasting approach may have difficulties with bias and properly calibrated uncertainty if used in a real-time setting. This includes rapidly changing dynamics [23, 60], both of the delay distribution and the underlying epidemic, retrospective changes in data [39] and long delays with few observed cases [41].

To avoid the aforementioned hindsight bias one can make their predictions publicly available in real-time [6, 43]. For the hospitalisations in Germany, Thomas Hotz and I have participated in the German COVID-19 NowcastHub [42] since November 2021 where nowcasts are available in a public Github repository [27] 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.

We use publicly available data from the German national health authority (RKI) on daily reported COVID-19 cases [47] and weekly reported hospitalisations [46]. 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.
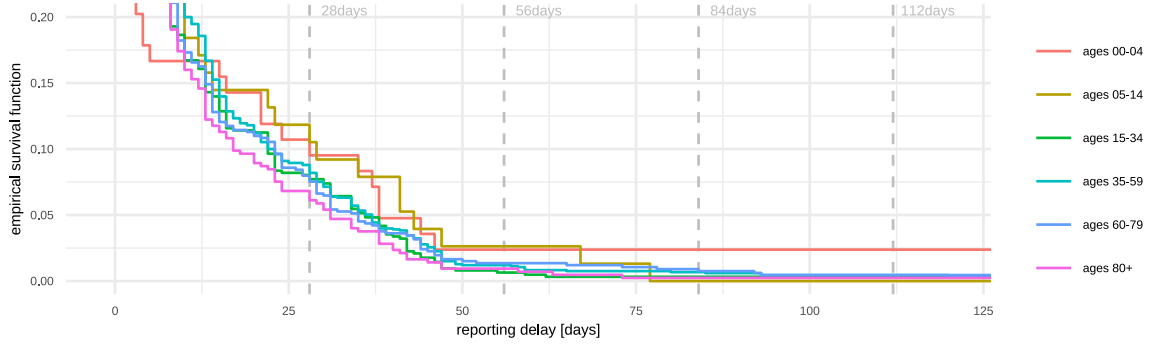
Figure 4.3: Survival function of reporting delays of weekly hospitalisations $H_{t,d}$ with case reporting date 01 September 2021. The delay distribution has long tails with a non-neglibgible fraction of observed delays longer than eight weeks, in some age groups even twelve weeks.

In line with the structure of the data provided by the RKI we let $H_{t,d}^a$ be the number of weekly hospitalisations in age group a with case reporting date $t-1, \ldots, t-7$ that are known on day $t+d$, aggregating over all states. Accordingly we define $I_{t,d}^a$ to be the number of weekly incident cases in age group a with reporting date $t-1, \ldots, t-7$ that are known on day $t+d$. Finally we reconstruct the reporting triangles for weekly hospitalisations (Figure 4.4) by differencing the $H_{t,d}^a$ for fixed t: $h_{t,d}^a = H_{t,d}^a - H_{t,d-1}^a$, setting $H_{t,-1}^a$ to 0 by convention. We recover the reporting triangle $i_{t,d}^a$ 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 $d = 0, \ldots, 28$.

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 $H_{t,d}$.

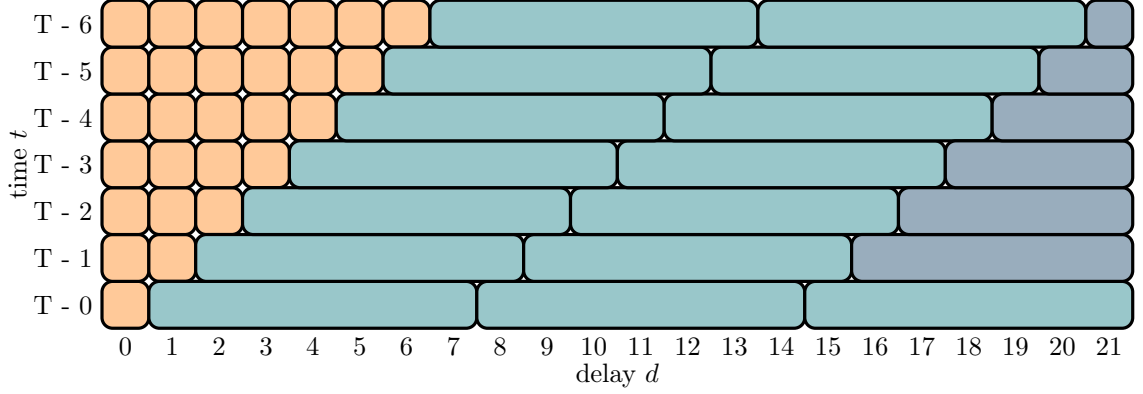Assuming that the proportions $p_{t,d,k}$ change slowly over time t we estimate them by

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.

$$\widehat{p_{t,d,k}} = \frac{H_{t-7k,d+7k} - H_{t-7k,d+7\cdot(k-1)}}{I_{t-7k,d}} = p_{t-7k,d,k} \tag{4.1}$$

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 $\eta_{t,s} = \log\left(H_{t,D} - H_{t,s-t}\right) - \log\left(\hat{H}_{t,D,s} - H_{t,s-t}\right)$. 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 slowly over time, especially in the older age groups. Due to smaller numbers of infections and hospitalisations reported in the younger age groups these fractions vary more strongly, occasionally 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
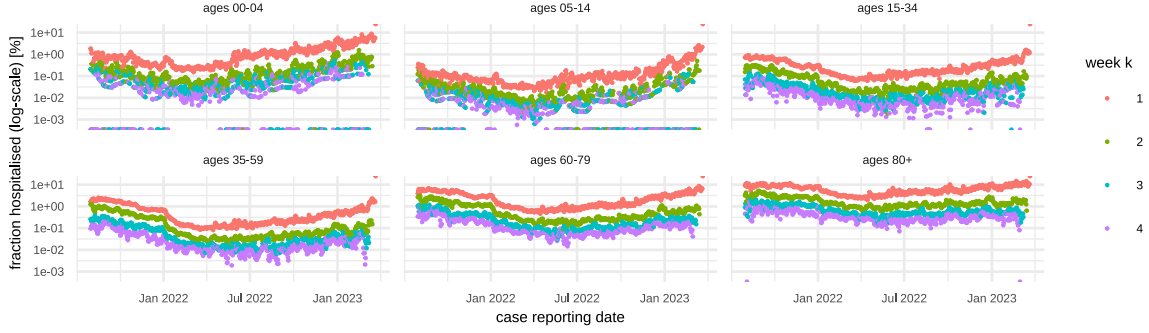
Figure 4.5: We show the fractions of hospitalisations in the k-th week after case reporting date t of initially reported cases in different age groups, i.e. $p_{t,0,k} = \left(H_{t,7k} - H_{t,7\cdot(k-1)}\right)/I_{t,0}$. Note the log-scale of the y-axis. During periods of low incidence, e.g. July – September, we find large fluctuations, 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.

variant became dominant in Germany [48]. 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 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 ≈ 10 days [18], 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.
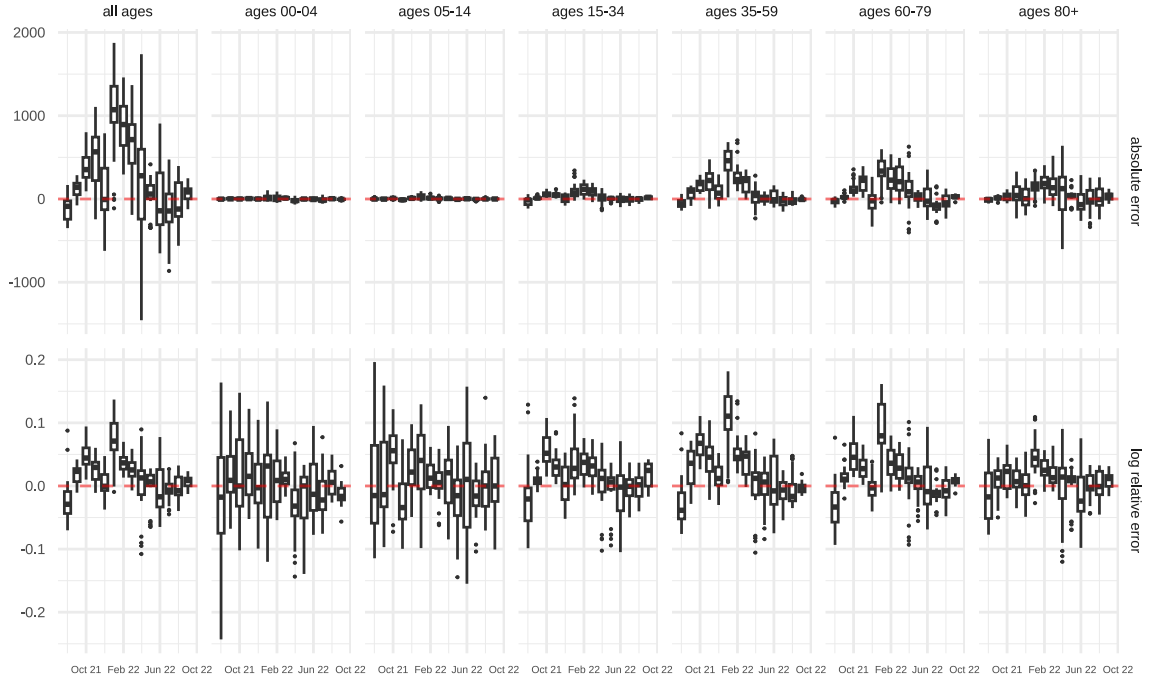
Figure 4.6: We show relative and absolute errors of prediction of our model for same day nowcasts by month of forecast and selected age groups. Relative errors are displayed on the log10 scale, i.e. as $\log_{10}(\text{predicted}) - \log_{10}(\text{actual})$. Up to December 2021 the model performs well, 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.
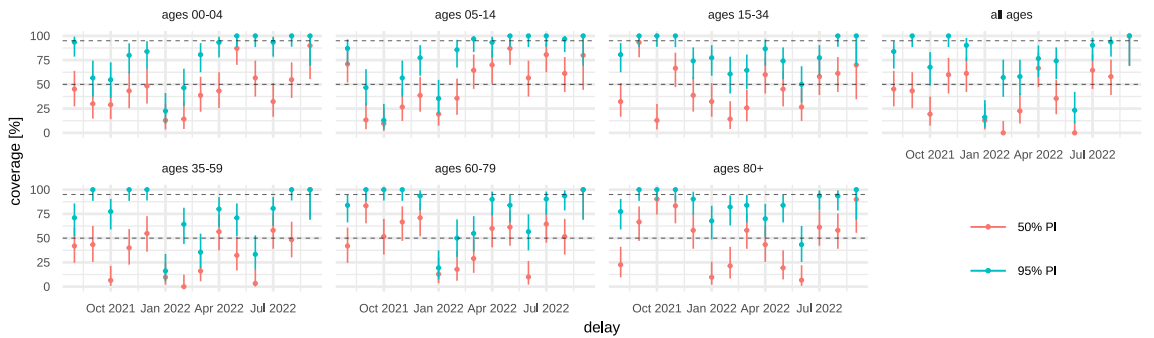
Figure 4.7: Empirical coverage of 50% and 95% prediction intervals (PI) based on same-day nowcasts for dates 2021-08-01 to 2022-09-10 (406 dates) for which the true amount of hospitalisations after 12 weeks is known as of the writing of this paper. We also display pointwise 95% binomial confidence intervals for the coverage. Given the difficulties of real-time forecasting [13] 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.

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: 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 [13], in particular quantifying uncertainty [6]. To test our model under real-time circumstances we submitted daily nowcasts to the German COVID-19 NowcastHub [42] since Novemer 2021. In the nowcasting context, [37] 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 hospitalisa-

tions 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 [42], 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 [54, 60], 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

Proofs (?)

Bibliography

[1] S. Agapiou et al. *Importance Sampling: Intrinsic Dimension and Computational Cost*. Jan. 14, 2017. DOI: `10.48550/arXiv.1511.06196`. arXiv: `1511.06196 [stat]`. preprint.

[2] Andrei R. Akhmetzhanov. "Estimation of Delay-Adjusted All-Cause Excess Mortality in the USA: March-December 2020." In: *Epidemiology and Infection* (2021). DOI: `10.1017/s0950268821001527`. pmid: `34210370`.

[3] Matthias An Der Heiden and Osamah Hamouda. "Schätzung Der Aktuellen Entwicklung Der SARS-CoV-2- Epidemie in Deutschland – Nowcasting." In: *Epidemiologisches Bulletin* (Apr. 22, 2020). DOI: `10.25646/6692.4`.

[4] Richard Arratia, Larry Goldstein, and Louis Gordon. "Poisson Approximation and the Chen-Stein Method." In: *Statistical Science* 5.4 (1990), pp. 403–424. ISSN: 0883-4237. JSTOR: `2245366`. URL: `https://www.jstor.org/stable/2245366` (visited on 01/11/2024).

[5] Leonardo Soares Bastos et al. "A Modelling Approach for Correcting Reporting Delays in Disease Surveillance Data." In: *Statistics in Medicine* (2019). DOI: `10.1002/sim.8303`.

[6] J. Bracher et al. "A Pre-Registered Short-Term Forecasting Study of COVID-19 in Germany and Poland during the Second Wave." In: *Nature Communications* 12.1 (1 Aug. 27, 2021), p. 5173. ISSN: 2041-1723. DOI: `10.1038/s41467-021-25207-0`. URL: `https://www.nature.com/articles/s41467-021-25207-0` (visited on 09/30/2021).

[7] Johannes Bracher et al. "National and Subnational Short-Term Forecasting of COVID-19 in Germany and Poland during Early 2021." In: *Communications Medicine* 2.1 (1 Oct. 31, 2022), pp. 1–17. ISSN: 2730-664X. DOI: `10.1038/s43856-022-00191-8`. URL: `https://www.nature.com/articles/s43856-022-00191-8` (visited on 11/16/2022).

[8] Jan M. Brauner et al. "Inferring the Effectiveness of Government Interventions against COVID-19." In: *Science* 371.6531 (Feb. 19, 2021), eabd9338. ISSN: 0036-8075, 1095-9203. DOI: `10.1126/science.abd9338`.

[9] Lawrence D. Brown. *Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory*. Lecture Notes-Monograph Series v. 9. Hayward, Calif: Institute of Mathematical Statistics, 1986. 283 pp. ISBN: 978-0-940600-10-2.

[10] Stephen Chan et al. "Count Regression Models for COVID-19." In: *Physica A: Statistical Mechanics and its Applications* 563 (Feb. 1, 2021), p. 125460. ISSN: 0378-4371. DOI: `10.1016/j.physa.2020.125460`.

[11] Sourav Chatterjee and Persi Diaconis. "The Sample Size Required in Importance Sampling." In: *Ann. Appl. Probab.* 28.2 (Apr. 1, 2018). ISSN: 1050-5164. DOI: `10.1214/17-AAP1326`.

[12] Nicolas Chopin and Omiros Papaspiliopoulos. *An Introduction to Sequential Monte Carlo*. Springer Series in Statistics. Cham, Switzerland: Springer, 2020. 378 pp. ISBN: 978-3-030-47844-5.

[13] Angel N. Desai et al. "Real-Time Epidemic Forecasting: Challenges and Opportunities." In: *Health Security* 17.4 (Aug. 2019), pp. 268–275. ISSN: 2326-5094. DOI: `10.1089/hs.2019.0022`.

[14] J. Durbin and S. J. Koopman. *Time Series Analysis by State Space Methods*. 2nd ed. Oxford Statistical Science Series 38. Oxford: Oxford University Press, 2012. 346 pp. ISBN: 978-0-19-964117-8.

[15] James Durbin and Siem Jan Koopman. "Monte Carlo Maximum Likelihood Estimation for Non-Gaussian State Space Models." In: *Biometrika* 84.3 (Sept. 1, 1997), pp. 669–684. ISSN: 0006-3444. DOI: `10.1093/biomet/84.3.669`.

[16] Max Ehre et al. "Certified Dimension Reduction for Bayesian Updating with the Cross-Entropy Method." In: *SIAM/ASA J. Uncertainty Quantification* 11.1 (Mar. 31, 2023), pp. 358–388. DOI: 10.1137/22M1484031.

[17] Michael Engel et al. "Bayesian Updating and Marginal Likelihood Estimation by Cross Entropy Based Importance Sampling." In: *Journal of Computational Physics* 473 (Jan. 15, 2023), p. 111746. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2022.111746.

[18] Christel Faes et al. "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 Jan. 2020), p. 7560. ISSN: 1660-4601. DOI: 10.3390/ijerph17207560.

[19] C. P. Farrington et al. "A Statistical Algorithm for the Early Detection of Outbreaks of Infectious Disease." In: *Journal of The Royal Statistical Society Series A-statistics in Society* (1996). DOI: 10.2307/2983331.

[20] Seth Flaxman et al. "Estimating the Effects of Non-Pharmaceutical Interventions on COVID-19 in Europe." In: *Nature* 584.7820 (Aug. 2020), pp. 257–261. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2405-7. pmid: 32512579.

[21] Sylvia Frühwirth-Schnatter. "Data Augmentation and Dynamic Linear Models." In: *Journal of Time Series Analysis* 15.2 (1994), pp. 183–202. ISSN: 1467-9892. DOI: 10.1111/j.1467-9892.1994.tb00184.x.

[22] "Discrete Spatial Variation." In: *Handbook of Spatial Statistics*. Ed. by Alan E. Gelfand et al. CRC Press, 2010. ISBN: 978-0-429-13650-4.

[23] Felix Günther et al. "Nowcasting the COVID-19 Pandemic in Bavaria." In: *Biometrical Journal* 63.3 (2021), pp. 490–502. ISSN: 1521-4036. DOI: 10.1002/bimj.202000112.

[24] Stefan Heyder and Thomas Hotz. "Measures of COVID-19 Spread." In: *Covid-19 pandisziplinär und international: Gesundheitswissenschaftliche, gesellschaftspolitische und philosophische Hintergründe*. Ed. by Alexander Kraemer and Michael Medzech. Medizin, Kultur, Gesellschaft. Wiesbaden: Springer Fachmedien, 2023, pp. 51–66. ISBN: 978-3-658-40525-0. DOI: 10.1007/978-3-658-40525-0_3. URL: https://doi.org/10.1007/978-3-658-40525-0_3 (visited on 10/21/2023).

[25] Michael Höhle and Matthias An Der Heiden. "Bayesian Nowcasting during the STEC O104:H4 Outbreak in Germany, 2011." In: *Biometrics* 70.4 (2014), pp. 993–1002. ISSN: 1541-0420. DOI: 10.1111/biom.12194.

[26] Tito Homem-de-Mello. "A Study on the Cross-Entropy Method for Rare-Event Probability Estimation." In: *INFORMS Journal on Computing* (July 20, 2007). DOI: 10.1287/ijoc.1060.0176.

[27] *Hospitalization Nowcast Hub*. KITmetricslab, Oct. 31, 2022. URL: https://github.com/KITmetricslab/hospitalization-nowcast-hub (visited on 11/09/2022).

[28] Norman Lloyd Johnson, Samuel Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*. 2nd ed. Wiley Series in Probability and Mathematical Statistics. New York: Wiley, 1994. 2 pp. ISBN: 978-0-471-58495-7 978-0-471-58494-0.

[29] Borus Jungbacker and Siem Jan Koopman. "Monte Carlo Estimation for Nonlinear Non-Gaussian State Space Models." In: *Biometrika* 94.4 (Dec. 1, 2007), pp. 827–839. ISSN: 0006-3444. DOI: 10.1093/biomet/asm074.

[30] Kenneth S. Kaminsky. "Prediction of IBNR Claim Counts by Modelling the Distribution of Report Lags." In: *Insurance Mathematics & Economics* 6.2 (Apr. 1, 1987), pp. 151–159. DOI: 10.1016/0167-6687(87)90024-2.

[31] H. J. Kappen and H. C. Ruiz. "Adaptive Importance Sampling for Control and Inference." In: *J Stat Phys* 162.5 (Mar. 1, 2016), pp. 1244–1266. ISSN: 1572-9613. DOI: 10.1007/s10955-016-1446-7.

[32] Yeganeh Khazaei et al. "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 (Nov. 2, 2023), p. 18900. ISSN: 2045-2322. DOI: 10.1038/s41598-023-45950-2.

[33] Augustine Kong, Jun S. Liu, and Wing Hung Wong. "Sequential Imputations and Bayesian Missing Data Problems." In: *Journal of the American Statistical Association* 89.425 (Mar. 1994), pp. 278–288. ISSN: 0162-1459. DOI: 10.1080/01621459.1994.10476469.

[34] Siem Jan Koopman, Rutger Lit, and Thuy Minh Nguyen. "Modified Efficient Importance Sampling for Partially Non-Gaussian State Space Models." In: *Statistica Neerlandica* 73.1 (2019), pp. 44–62. ISSN: 1467-9574. DOI: 10.1111/stan.12128.

[35] Pierre Simon Laplace. "Memoir on the Probability of the Causes of Events." In: *Statistical Science* 1.3 (Aug. 1986), pp. 364–378. ISSN: 0883-4237, 2168-8745. DOI: 10.1214/ss/1177013621.

[36] Steffen L. Lauritzen. *Graphical Models*. Oxford Statistical Science Series 17. Oxford : New York: Clarendon Press ; Oxford University Press, 1996. 298 pp. ISBN: 978-0-19-852219-5.

[37] J. F. Lawless. "Adjustments for Reporting Delays and the Prediction of Occurred but Not Reported Events." In: *Canadian Journal of Statistics* 22.1 (1994), pp. 15–31. ISSN: 1708-945X. DOI: 10.2307/3315826.n1.

[38] Sarah F. McGough et al. "Nowcasting by Bayesian Smoothing: A Flexible, Generalizable Model for Real-Time Epidemic Tracking." In: *PLOS Computational Biology* (2020). DOI: 10.1371/journal.pcbi.1007735. pmid: 32251464.

[39] Douglas N. Midthune et al. "Modeling Reporting Delays and Reporting Corrections in Cancer Registry Data." In: *Journal of the American Statistical Association* 100.469 (Mar. 1, 2005), pp. 61–70. ISSN: 0162-1459. DOI: 10.1198/016214504000001899.

[40] M. Morf and T. Kailath. "Square-Root Algorithms for Least-Squares Estimation." In: *IEEE Transactions on Automatic Control* 20.4 (Aug. 1975), pp. 487–497. ISSN: 1558-2523. DOI: 10.1109/TAC.1975.1100994.

[41] Angela Noufaily et al. "Modelling Reporting Delays for Outbreak Detection in Infectious Disease Data." In: *Journal of The Royal Statistical Society Series A-statistics in Society* (2015). DOI: 10.1111/rssa.12055.

[42] *Nowcasts Der COVID-19 Hospitalisierungsinzidenz*. 2022. URL: https://covid19nowcasthub.de/ (visited on 11/09/2022).

[43] Evan L. Ray et al. "Ensemble Forecasts of Coronavirus Disease 2019 (COVID-19) in the U.S." In: *medRxiv* (Aug. 22, 2020), p. 2020.08.19.20177493. DOI: 10.1101/2020.08.19.20177493.

[44] Arthur E. Renshaw and Richard J. Verrall. "A Stochastic Model Underlying the Chain-Ladder Technique." In: *British Actuarial Journal* 4.4 (1998), pp. 903–923. DOI: 10.1017/S1357321700000222.

[45] Jean-Francois Richard and Wei Zhang. "Efficient High-Dimensional Importance Sampling." In: *Journal of Econometrics* 141.2 (Dec. 1, 2007), pp. 1385–1411. ISSN: 0304-4076. DOI: 10.1016/j.jeconom.2007.02.007.

[46] Robert Koch-Institut. *COVID-19-Hospitalisierungen in Deutschland*. Version 2021-10-01. Zenodo, Oct. 1, 2021. DOI: 10.5281/ZENODO.5519056.

[47] Robert Koch-Institut. *SARS-CoV-2 Infektionen in Deutschland*. Version 2022-02-07. Zenodo, Feb. 7, 2022. DOI: 10.5281/ZENODO.4681153.

[48] Robert Koch-Institut. *Wöchentlicher Lagebericht Des RKI Zur Coronavirus-Krankheit-2019 (COVID-19)*. Jan. 2022.

[49] Nathan Ross. "Fundamentals of Stein's Method." In: *Probab. Surveys* 8 (none Jan. 1, 2011). ISSN: 1549-5787. DOI: 10.1214/11-PS182.

[50] Reuven Rubinstein. "The Cross-Entropy Method for Combinatorial and Continuous Optimization." In: *Methodology and Computing in Applied Probability* 1.2 (Sept. 1, 1999), pp. 127–190. ISSN: 1573-7713. DOI: 10.1023/A:1010091220143.

[51] Reuven Y. Rubinstein and Dirk P. Kroese. *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. New York, NY: Springer New York, 2004. ISBN: 978-1-4757-4321-0.

[52] Maëlle Salmon et al. "Bayesian Outbreak Detection in the Presence of Reporting Delays." In: *Biometrical Journal* (2015). DOI: 10.1002/bimj.201400159. pmid: 26250543.

[53] Florian Schäfer, Matthias Katzfuss, and Houman Owhadi. "Sparse Cholesky Factorization by Kullback–Leibler Minimization." In: *SIAM J. Sci. Comput.* 43.3 (Jan. 2021), A2019–A2046. ISSN: 1064-8275. DOI: 10.1137/20M1336254.

[54] Marc Schneble et al. "Nowcasting Fatal COVID-19 Infections on a Regional Level in Germany." In: *Biometrical Journal* 63.3 (Mar. 2021), pp. 471–489. ISSN: 0323-3847, 1521-4036. DOI: 10.1002/bimj.202000143.

[55] Wolfgang Schneider. *Der Kalmanfilter Als Instrument Zur Diagnose Und Schätzung Variabler Parameter in Ökonometrischen Modellen*. Arbeiten Zur Angewandten Statistik Bd. 27. Heidelberg: Physica-Verlag, 1986. 490 pp. ISBN: 978-3-7908-0359-4.

[56] Neil Shephard and Michael K. Pitt. "Likelihood Analysis of Non-Gaussian Measurement Time Series." In: *Biometrika* 84.3 (Sept. 1, 1997), pp. 653–667. ISSN: 0006-3444. DOI: 10.1093/biomet/84.3.653.

[57] K. Sherratt et al. *Predictive Performance of Multi-Model Ensemble Forecasts of COVID-19 across European Nations*. June 16, 2022. DOI: 10.1101/2022.06.16.22276024. URL: http://medrxiv.org/lookup/doi/10.1101/2022.06.16.22276024 (visited on 11/28/2022). preprint.

[58] Luke Tierney and Joseph B. Kadane. "Accurate Approximations for Posterior Moments and Marginal Densities." In: *Journal of the American Statistical Association* 81.393 (Mar. 1986), pp. 82–86. ISSN: 0162-1459, 1537-274X. DOI: 10.1080/01621459.1986.10478240.

[59] Luke Tierney, Robert E. Kass, and Joseph B. Kadane. "Fully Exponential Laplace Approximations to Expectations and Variances of Nonpositive Functions." In: *Journal of the American Statistical Association* 84.407 (Sept. 1989), pp. 710–716. ISSN: 0162-1459, 1537-274X. DOI: 10.1080/01621459.1989.10478824.

[60] Jan van de Kassteele et al. "Nowcasting the Number of New Symptomatic Cases during Infectious Disease Outbreaks Using Constrained P-Spline Smoothing." In: *Epidemiology* (2019). DOI: 10.1097/ede.0000000000001050. pmid: 31205290.

[61] Aad W. Van der Vaart. *Asymptotic Statistics*. Cambridge: Cambridge University Press, 2000.

[62] Halbert White. "Maximum Likelihood Estimation of Misspecified Models." In: *Econometrica* 50.1 (1982), pp. 1–25. ISSN: 0012-9682. DOI: 10.2307/1912526. JSTOR: 1912526.

[63] Scott L. Zeger, Lai-Chu See, and Peter J. Diggle. "Statistical Methods for Monitoring the AIDS Epidemic." In: *Statistics in Medicine* 8.1 (1989), pp. 3–21. DOI: 10.1002/sim.4780080104.

[64] Wei Zhang et al. "Applications of the Cross-Entropy Method to Importance Sampling and Optimal Control of Diffusions." In: *SIAM J. Sci. Comput.* 36.6 (Jan. 2014), A2654–A2672. ISSN: 1064-8275. DOI: 10.1137/14096493X.

Declaration

Put your declaration here.

Ilmenau, October 2023