STATE SPACE MODELS FOR REGIONAL EPIDEMIOLGICAL INDICATORS

STEFAN HEYDER

October 2023 – Draft v 0.1

ABSTRACT

Short summary of the contents in English…

ZUSAMMENFASSUNG

Kurze Zusammenfassung des Inhaltes in deutscher Sprache…

PUBLICATIONS AND CONTRIBUTIONS

This thesis consists of mostly unpublished work.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

[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, Stefan Heyder, Thomas Hotz, and Tyll Krueger. "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 M. Grundel, Stefan Heyder, Thomas Hotz, Tobias K. S. Ritschel, Philipp Sauerteig, and Karl Worthmann. "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).

[5] Sara Grundel, Stefan Heyder, Thomas Hotz, Tobias K. S. Ritschel, Philipp Sauerteig, and Karl Worthmann. "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).

[6] Thomas Hotz, Matthias Glock, Stefan Heyder, Sebastian Semper, Anne Böhle, and Alexander Krämer. "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

1 INTRODUCTION

2 EPIDEMIOLOGICAL DATA & MODELS

3 STATE SPACE MODELS

State space models ...

Definition 3.1 (State Space Model)**.** A **state space model** is a discrete time stochastic process $(X_t, Y_t)_{t=0,\dots,n-1}$ taking values in $\mathcal{X} \times \mathcal{Y}$ such that

1. The marginal distribution of the **states** $(X_0, \dots, X_{n-1})$ is a discrete time Markov process, i.e. if $s < t$ then X_t is conditionally independent of X_s given X_{t-1}.

2. Conditional on the state X_t and observation Y_{t-1}, Y_t is independent of X_s and Y_{s-1}, $s < t$.

For notational convenience we will write $\boldsymbol{X}_{s:t} = (X_s, \dots, X_t)$ for the vector that contains all states from s to t, dropping the index $s : t$ if we consider the whole set of observations, so $\boldsymbol{X} = \boldsymbol{X}_{0:n-1}$ Similarly we set $\boldsymbol{Y}_{s:t} = (Y_s, \dots, Y_t)$ and $\boldsymbol{Y} = \boldsymbol{Y}_{0:n-1}$.

picutre of dependency structure

Remark. Contrary to the standard definition of a state space model, our Definition 3.1 allows Y_t to depend on Y_{t-1}. This is not a limitation of the standard definition: given a state space model 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 state space model becomes a stochastic process on $(\mathcal{X} \times \mathcal{Y}) \times \mathcal{Y}$. As Chapter 4 will make extensive use of state space models with this dependency structure we opt to use this non-standard definition here.

In most models I consider in this thesis I will 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.

Given data $(y_t)_{t=0,\dots,n-1}$ that may be modeled with a state space model the practitioner is confronted with several tasks:

1. Choosing a suitable, usually parametric, class of state space models 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 $(X_0, \dots, X_{n-1})$ from the observations by determining, either analytically or through simulation, the **smoothing distribution** $\boldsymbol{X}|\boldsymbol{Y}$ or various functionals thereof.

The first step, item 1, requires that the practitioner to specfiy a joint probability distribution for the states and observations. Due to the assumed dependency structure this boils down to specifying transition kernels for the states and observations .

ausführlicher

Most models that I 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) domination measure $\mu_{\mathcal{Y}}$.

check whether there a models that violate th

Notation (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 $\boldsymbol{X}$, $p(x_t|x_{t-1})$ the conditional density of $X_t|X_{t-1}$ and similaryl for observations: $p(y|x)$ is the density of all observations $\boldsymbol{Y}$ conditional on all states $\boldsymbol{X}$.

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 $\theta \in \Theta$ and the dependence of the model on θ is of interest I 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 state space model factorizes as

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

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

Refer to Chapter 4 for constructing parametric state space models that capture relevant epidemiological effects.

Regarding the second step, item 2, a frequentist practitioner will want to perform maximum likelihood inference on θ. While confidence intervals for θ can be derived both theoretically and practically, they are usually of little interest . cite chopin?

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

$$
p(y) = \int_{\mathcal{X}^n} p(x,y)\,\mathrm{d}\,x, \tag{3.1}
$$

which is not analytically available except in special models (see Section 3.1). Direct numerical evaluation of Equation (3.1) is usually hopeless due to the high dimensionality of the state space $\mathcal{X}^n$. Instead one resorts to simulation based inference by importance sampling (see Section 3.2) or particle filters . cite something

3.1 LINEAR GAUSSIAN STATE SPACE MODELS

3.2 IMPORTANCE SAMPLING FOR STATE SPACE MODELS

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 $X^i \overset{\text{i.i.id}}{\sim} p(x)$. 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 are able to generate (ap- cite proximate) 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 $\ell(\theta)$.

In more general terms, let $g : \mathcal{X} \to \mathbf{R}$ be a function whose integral

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

we want to compute. Furthermore suppose that we can write

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

for a probability measure $\mathbf{P}$ and function $f : \mathcal{X} \to \mathbf{R}$. Let $\mathbf{Q}$ be another measure on $\mathcal{X}$ such that $\mathbf{P}$ is absolutely continuous w.r.t. $\mathbf{Q}$, $\mathbf{P} \ll Q$.

4 STATE SPACE MODELS FOR EPIDEMIOLOGICAL DATA

5 CONCLUSION

A

APPENDIX 01

A.1 SOME APPENDIX SECTION

DECLARATION

Put your declaration here.

Ilmenau, October 2023

Stefan Heyder