

---

# Appendix: A Probabilistic State Space Model for Joint Inference from Differential Equations and Data

---

**Jonathan Schmidt**  
University of Tübingen  
Tübingen, Germany

jonathan.schmidt@uni-tuebingen.de

**Nicholas Krämer**  
University of Tübingen  
Tübingen, Germany

nicholas.kraemer@uni-tuebingen.de

**Philipp Hennig**  
University of Tübingen  
Max Planck Institute for Intelligent Systems  
Tübingen, Germany  
philipp.hennig@uni-tuebingen.de

## A Implementation details

This section provides detailed information about the state-space model and approximate Gaussian inference therein. Appendix A.1 defines the augmented state-space model that formalizes the dynamics of the Gauss–Markov processes introduced in Section 3.1. Appendix A.2 provides the equations for prediction and update steps of the extended Kalman filter in such a setup, which is described in Section 3.4 (in particular, Algorithm 1).

### A.1 Augmented state-space model

Section 3 describes the joint inference of both a latent process  $u(t) : [t_0, t_{\max}] \rightarrow \mathbb{R}^l$  that parametrizes an ODE and  $x(t) : [t_0, t_{\max}] \rightarrow \mathbb{R}^d$ , the solution of said ODE. The dynamics of the processes are modeled by the stochastic differential equation

$$d \begin{pmatrix} U(t) \\ X(t) \end{pmatrix} = \underbrace{\begin{pmatrix} F_U & 0 \\ 0 & F_X \end{pmatrix}}_{=:F} \begin{pmatrix} U(t) \\ X(t) \end{pmatrix} dt + \underbrace{\begin{pmatrix} L_U & 0 \\ 0 & L_X \end{pmatrix}}_{=:L} d \begin{pmatrix} W_U(t) \\ W_X(t) \end{pmatrix}, \quad (\text{A.1})$$

with Gaussian initial conditions

$$\begin{pmatrix} U(t_0) \\ X(t_0) \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} m_U(t_0) \\ m_X(t_0) \end{pmatrix}, \begin{pmatrix} P_U(t_0) & 0 \\ 0 & P_X(t_0) \end{pmatrix} \right). \quad (\text{A.2})$$

The block-diagonal structure is due to the independent dynamics of the prior processes. The *drift matrices*  $F_U$  and  $F_X$ , as well as the *dispersion matrices*  $L_U$  and  $L_X$  depend on the choice of the respective processes  $U$  and  $X$ . The measurement models are given in Eq. (6) (for observed data) and in Eq. (7) (for ODE measurements).

In the experiments presented in Sections 5.2 and 5.3 we model the latent contact rate  $\beta(t)$  as a Matérn- $3/2$  process with characteristic length scale  $\ell_q$ . Hence,

$$dU(t) = \underbrace{\begin{pmatrix} 0 & 1 \\ -(\sqrt{3}/\ell_q)^2 & -2\sqrt{3}/\ell_q \end{pmatrix}}_{F_U} U(t) dt + \underbrace{\begin{pmatrix} 0 \\ 1 \end{pmatrix}}_{L_U} dW_U(t). \quad (\text{A.3})$$

The SIRD counts are modeled as the twice-integrated Wiener process

$$dX(t) = \underbrace{\begin{pmatrix} 0 & I_d & 0 \\ 0 & 0 & I_d \\ 0 & 0 & 0 \end{pmatrix}}_{F_X} X(t) dt + \underbrace{\begin{pmatrix} 0 \\ 0 \\ I_d \end{pmatrix}}_{L_X} dW_X(t), \quad (\text{A.4})$$

such that  $X = (X^{(0)}, X^{(1)}, X^{(2)})^\top$  models the SIRD counts and the first two derivatives. Notice that  $F_X \in \mathbb{R}^{d(\nu+1) \times d(\nu+1)}$  and  $L_X \in \mathbb{R}^{d(\nu+1) \times d}$  are block matrices.  $I_d$  denotes the  $d \times d$  identity matrix. In the context of the experiments,  $d = 4$  (S, I, R, and D) and  $\nu = 2$  (twice-integrated Wiener process). More details on the use of integrated Wiener processes in probabilistic ODE solvers can be found in, for instance, the work by Kersting et al. [5].

## A.2 Kalman filter equations

This section is concerned with the exact steps that make up the algorithm summarized in Section 3.4. The stochastic differential equation defined in Eq. (A.1) formalizes the dynamics of the processes  $U(t)$  and  $X(t)$  that model  $u(t)$  and  $x(t)$ , respectively. Define  $\Delta t := t_j - t_{j-1} > 0$  for all  $t_j = t_1, \dots, t_{\max}$ . The *transition densities* of  $U$  and  $X$  are [3]

$$U(t + \Delta t) | U(t) \sim \mathcal{N}(\Phi_U(\Delta t)U(t), Q_U(\Delta t)), \quad (\text{A.5a})$$

$$X(t + \Delta t) | X(t) \sim \mathcal{N}(\Phi_X(\Delta t)X(t), Q_X(\Delta t)), \quad (\text{A.5b})$$

where transition matrices  $\Phi_U(\Delta t) \in \mathbb{R}^{\ell \times \ell}$  and  $\Phi_X(\Delta t) \in \mathbb{R}^{d(\nu+1) \times d(\nu+1)}$ , as well as the process noise covariances  $Q_U(\Delta t) \in \mathbb{R}^{\ell \times \ell}$  and  $Q_X(\Delta t) \in \mathbb{R}^{d(\nu+1) \times d(\nu+1)}$  are available in closed form and can be computed, for instance, with matrix fraction decomposition [12, 1].

Define the transition matrix and process noise covariance of the process in Eq. (A.1) as

$$\Phi(\Delta t) := \begin{pmatrix} \Phi_U(\Delta t) & 0 \\ 0 & \Phi_X(\Delta t) \end{pmatrix}, \quad Q(\Delta t) := \begin{pmatrix} Q_U(\Delta t) & 0 \\ 0 & Q_X(\Delta t) \end{pmatrix}. \quad (\text{A.6})$$

Further, let

$$\begin{pmatrix} U(t_j) \\ X(t_j) \end{pmatrix} \sim \mathcal{N}(m_j, P_j), \quad (\text{A.7})$$

for time points  $t_j \in \mathcal{T} = \mathcal{T}^{\text{OBS}} \cup \mathcal{T}^{\text{ODE}}$ . The predicted mean and covariance  $m_j^-$  and  $P_j^-$  are

$$m_j^- = \Phi(\Delta t) m_{j-1}, \quad (\text{A.8})$$

$$P_j^- = \Phi(\Delta t) P_{j-1} \Phi(\Delta t)^\top + Q(\Delta t), \quad (\text{A.9})$$

for given initial conditions  $m_0, P_0$ . The prediction step is the same, for both  $t_j \in \mathcal{T}^{\text{OBS}}$  and  $t_j \in \mathcal{T}^{\text{ODE}}$ .

As detailed in Section 3, two different update steps are defined for two kinds of observations. When observing data  $y_{0:N}$ , i.e.  $t_n \in \mathcal{T}^{\text{OBS}}$ , the update step follows the rules of a standard Kalman filter. The updated mean  $m_n$  and covariance  $P_n$  at time  $t_n$  are computed as

$$v_n = y_n - H m_n^-, \quad (\text{A.10})$$

$$S_n = H P_n^- H^\top + R, \quad (\text{A.11})$$

$$K_n = P_n^- H^\top S_n^{-1}, \quad (\text{A.12})$$

$$m_n = m_n^- + K_n v_n, \quad (\text{A.13})$$

$$P_n = P_n^- - K_n S_n K_n^\top. \quad (\text{A.14})$$

The matrices  $H$  and  $R$  are defined as in Eq. (6) in the paper.

Recall the ODE measurement model from Eq. (7), which we here denote as  $h$ , as

$$h \left( \begin{pmatrix} U(t) \\ X(t) \end{pmatrix} \right) = X^{(1)} - f(X^{(0)}; U(t)). \quad (\text{A.15})$$

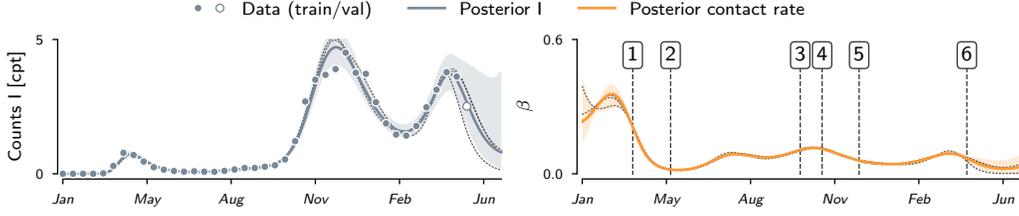


Figure 1: **Estimated counts of infectious cases and contact rate.** The estimates are obtained from MCMC sampling in an SIRD model with a parametric function for the contact rate  $\beta(t)$ . The case counts of infectious people are scaled to cases per thousand (cpt). The shaded areas show the 95% credible interval and the dotted black lines are samples from the posterior. Compared to the non-parametric approach presented in the paper, the estimate over  $\beta(t)$  is very confident in general. The posterior mean closely resembles the results obtained in Sections 5.2 and 5.3. The numbered markers in the right plot are explained in Table 1 in the paper.

At locations  $t_m \in \mathcal{T}^{\text{ODE}}$ , we condition on the ODE measurements  $z_{0:M}$ . Recall that these pseudo-observations are all zero. According to Eq. (10.79) in the book by Särkkä and Solin [11],

$$v_m = z_m - h(m_m^-), \quad (\text{A.16})$$

$$S_m = [Dh(m_m^-)] P_m^- [Dh(m_m^-)]^\top + \lambda^2 I_d, \quad (\text{A.17})$$

$$K_m = P_m^- [Dh(m_m^-)]^\top S_m^{-1}, \quad (\text{A.18})$$

$$m_m = m_m^- + K_m v_m, \quad (\text{A.19})$$

$$P_m = P_m^- - K_m S_m K_m^\top, \quad (\text{A.20})$$

where  $[Dh(m_m^-)]$  denotes the Jacobian of  $h$  at  $m_m^-$ . In the case of a Dirac likelihood (see Eq. (7)),  $\lambda^2 = 0$  holds. For numerical stability (especially for  $\lambda^2 = 0$ ) one can instead implement square-root filtering (see, e.g., [3, 6]). All experiments in Section 5 use square-root filtering.

## B Parametric model for MCMC sampling

This section first introduces a functional form for  $\beta(t)$  that connects to the non-parametric model introduced in Section 3. Then, a generative model for Markov-chain Monte Carlo (MCMC) inference over the unknown parameters of  $\beta(t)$  is set up.

We establish a parametric model for the latent, time-varying contact rate in an SIRD model in terms of Fourier features. In light of Mercer’s theorem and the fact that stationary covariance functions have complex-exponential eigenfunctions [10, Chapter 4.3], this closely connects to the Matérn- $\frac{3}{2}$  process used in Sections 5.2 and 5.3 (see also [9]).

Concretely, we proceed as follows. Let  $\mathbb{T}$  denote a dense time grid. First, (i) compute the kernel Gram matrix  $K$  on  $\mathbb{T}$ , such that  $(K)_{ij} = k(x_i, x_j)$  with  $x_i, x_j \in \mathbb{T}$ .  $k$  is the Matérn- $\frac{3}{2}$  covariance function. As in the experiments before, we set the characteristic lengthscale to  $\ell = 75$ . Then, (ii) compute the eigendecomposition of  $K$ . In order to keep the dimensionality of the inference problem feasible, select  $r \ll |\mathbb{T}|$  eigenvectors that correspond to the  $r$  largest eigenvalues of  $K$ . In this experiment, we choose  $r = 25$ . (iii) For each eigenvector, the strongest frequency component  $\omega$  is determined by the discrete Fourier decomposition. This yields a set of frequencies  $\{\omega_i : i = 1, \dots, r\}$ . Finally, the parametric model is defined as the sum of parametrized Fourier features of the form

$$\beta(t) = \vartheta \left( \sum_{i=1}^r a_i \cos(2\pi\omega_i t) + b_i \sin(2\pi\omega_i t) \right), \quad (\text{B.1})$$

where  $\vartheta$  is the logistic sigmoid function as described in Section 5. We aim to compute a posterior contact rate  $\beta(t)$  by MCMC inference over the coefficients  $a_i$  and  $b_i$ ,  $i = 1, \dots, r$ . To this end, we define a prior over the parameter vector  $\theta := (a_1, b_1, \dots, a_r, b_r)^\top$  and a likelihood for the COVID-19 case counts  $y_{0:N}$  with respect to  $\theta$ .

In order to ensure non-negative case counts, as in Section 5.3, we assume log-normally distributed measurements with i.i.d. noise

$$p(y_{0:N} | \theta) = \prod_{n=0}^N \text{LogNormal} \left( y_n; \log \left( x^{(\theta)}(t_n) \right), \sigma^2 I_{2r} \right), \quad (\text{B.2})$$

where  $\sigma^2$  is inferred from the data along with  $\theta$ .  $x^{(\theta)}(t_n)$  denotes the solution of the SIRD system at time  $t_n$ , parametrized by the vector of coefficients  $\theta$  through the contact rate from Eq. (B.1). Notably, each evaluation of the likelihood involves numerically integrating the SIRD system, which significantly increases the computational cost entailed by the inference algorithm. This is done by NumPyro’s DOPRI-5 implementation [8, 2].

The prior distributions over the Fourier-feature coefficients and over  $\sigma^2$  are chosen as

$$p(\theta) = \mathcal{N}(\theta; \mu_\theta, \Sigma_\theta), \quad p(\sigma^2) = \text{HalfCauchy}(\sigma^2; 0.01). \quad (\text{B.3})$$

The mean  $\mu_\theta$  of the prior over  $\theta$  is set to a maximum-likelihood estimate by minimizing the negative logarithm of Eq. (B.2) with SciPy’s L-BFGS optimization algorithm [13, 7]. The covariance is chosen as  $\Sigma_\theta = 0.1 \cdot I_{2r}$ .

The goal of the experiment is to compute a posterior over the coefficients  $\theta$  (and the measurement covariance  $\sigma^2$ ) that is comparable to the results obtained in Sections 5.2 and 5.3. Like before, recovery rate and fatality rate are assumed fixed and known at  $\gamma = 0.06$  and  $\eta = 0.002$ . We compute the posterior  $p(\theta | y_{0:N})$  using NumPyro’s implementation of the No-U-Turn sampler [4].

Figure 1 shows the estimated number of infectious people and the contact rate over time as inferred by the MCMC algorithm. The state estimate matches the data points well and the uncertainty increases when extrapolating. Like in the experiments in Sections 5.2 and 5.3, the final 14 observations serve as a validation set and the model extrapolates 31 days into the future. The posterior mean closely resembles the results obtained from our method. However, the uncertainty is lower in general, especially in the beginning and over the summer months.

## C Sources for governmental measures in Germany

This section provides the sources used to list the governmental measures in Table 1. In order to provide reliable sources, we refer to the official press releases, as published by the German government. For each policy change, we provide a very brief idea of the imposed measures and official sources by the German government (only available in German language).

### C.1 March 22, 2020 (Mark 1)

Citizens are urged to restrict social contacts as much as possible and the formation of groups is sanctioned in public spaces as well as at home.

<https://www.bundesregierung.de/breg-de/themen/coronavirus/besprechung-der-bundeskanzlerin-mit-den-regierungschefinnen-und-regierungschefs-der-laender-vom-22-03-2020-1733248>

<https://www.bundesregierung.de/resource/blob/975226/1733246/e6d6ae0e89a7ffea1ebf6f32cf472736/2020-03-22-mpk-data.pdf?download=1>

### C.2 May 6, 2020 (Mark 2)

The government puts the federal states in charge of appropriately relaxing the imposed measures. Different states handle the situation differently, according to the respective incidences (*‘hotspot strategy’*).

<https://www.bundesregierung.de/breg-de/aktuelles/pressekonferenzen/pressekonferenz-von-bundeskanzlerin-merkel-ministerpraesident-soeder-und-dem-ersten-buergermeister-tschantcher-im-anschluss-an-das-gespraech-mit-den-regierungschefinnen-und-regierungschefs-der-laender-1751050>

### C.3 October 7, 2020 (Mark 3) and October 14, 2020

The population is again urged to restrict contacts if possible.

<https://www.bundeskanzlerin.de/bkin-de/aktuelles/telefonschaltkonferenz-des-chefs-des-bundeskanzleramts-mit-den-chefinnen-und-chefs-der-staats-und-senatskanzleien-der-laender-am-7-oktober-2020-1796770>

One week later, new light restrictions are imposed. The number of people allowed in social gatherings is limited, according to local incidences.

<https://www.bundesregierung.de/resource/blob/997532/1798920/9448da53f1fa442c24c37abc8b0b2048/2020-10-14-beschluss-mpk-data.pdf?download=1>

#### **C.4 November 2, 2020 (Mark 4)**

Partial shutdown of public life ('*lockdown light*'). Across the country, the number of people allowed in social gatherings is limited to ten, where the number of households present must not exceed two. Most of public services are closed or offered only virtually, if possible.

<https://www.bundesregierung.de/breg-de/aktuelles/videokonferenz-der-bundeskanzlerin-mit-den-regierungschefinnen-und-regierungschefs-der-laender-am-28-oktober-2020-1805248>

#### **C.5 December 16, 2020 (Mark 5)**

Across the country, the number of people allowed in social gatherings is limited to five, where the number of households present must not exceed two. Except for stores of systemic importance, the retail sector is mostly shut down.

<https://www.bundesregierung.de/resource/blob/997532/1827366/69441fb68435a7199b3d3a89bff2c0e6/2020-12-13-beschluss-mpk-data.pdf?download=1>

#### **C.6 April 23, 2021 (Mark 6)**

The aforementioned measures were mostly governed and implemented by the respective federal states. On April 22, 2021, the German government decides on a nationwide decree of measures to come into effect on the following day (April 23, 2021). Depending on the seven-day incidence, curfews, contact restrictions, and a shutdown of large parts of public life are imposed.

[https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/3\\_Downloads/Gesetze\\_und\\_Verordnungen/GuV/B/4\\_BevSchG\\_BGBL.pdf](https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/3_Downloads/Gesetze_und_Verordnungen/GuV/B/4_BevSchG_BGBL.pdf)

## **References**

- [1] P. Axelsson and F. Gustafsson. Discrete-time solutions to the continuous-time differential Lyapunov equation with applications to Kalman filtering. *IEEE Transactions on Automatic Control*, 60(3):632–643, 2015.
- [2] J. Dormand and P. Prince. A family of embedded Runge-Kutta formulae. *Journal of Computational and Applied Mathematics*, 6(1):19–26, 1980.
- [3] M. Grewal and A. Andrews. *Kalman Filtering: Theory and Practice Using MATLAB*. Wiley, 2011.
- [4] M. D. Hoffman and A. Gelman. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.
- [5] H. Kersting, T. J. Sullivan, and P. Hennig. Convergence rates of Gaussian ODE filters. *Statistics and Computing*, 30(6):1791–1816, 2020.
- [6] N. Krämer and P. Hennig. Stable implementation of probabilistic ODE solvers. *arXiv:2012.10106*, 2020.
- [7] D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1-3):503–528, 1989.
- [8] D. Phan, N. Pradhan, and M. Jankowiak. Composable effects for flexible and accelerated probabilistic programming in NumPyro. *arXiv:1912.11554*, 2019.

- [9] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, 2008.
- [10] C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.
- [11] S. Särkkä and A. Solin. *Applied Stochastic Differential Equations*. Cambridge University Press, 2019.
- [12] R. Stengel. *Optimal Control and Estimation*. Dover Publications, 1994.
- [13] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 2020.