State Space Gaussian Processes with Non-Gaussian Likelihood

Hannes Nickisch[1] Arno Solin[2] Alexander Grigorievskiy[2,3]

[1]Digital Imaging, Philips Research, Hamburg, Germany
[2]Department of Computer Science, Aalto University, Espoo, Finland
[3]Silo.AI, Helsinki, Finland

Introduction

- Overview and tooling for temporal Gaussian process (GP) modeling with non-Gaussian likelihoods.
- By reformulating the GP into a state space model, inference can be done in $\mathcal{O}(n)$ time complexity.
- Means of combining efficient state space methodology with approximate inference schemes for non-Gaussian likelihoods.
- Covered approximate inference algorithms:
  - Laplace Approximation (LA)
  - Variational Bayes (VB)
  - Direct KL minimization (KL)
  - Single-sweep Expectation Propagation (EP) / assumed density filtering (ADF)
- Code is available in the GPML toolbox v. 4.2.

Temporal Gaussian Processes

- GPs [2] are handy probabilistic tools for regression and classification
- Consider a dataset of input–output pairs: $\mathcal{D} = \{(t_i, y_i)\}_{i=1}^n$
- The GP model can be written as:

$$f(t) \sim \mathcal{GP}(m(t), k(t,t')) \qquad \text{GP prior}$$

$$\mathbf{y} \,|\, \mathbf{f} \sim \prod_{i=1}^{n} \mathbb{P}(y_i \,|\, f(t_i)) \qquad \text{Likelihood}$$

- The prior assumptions are encoded in the covariance function $k(\cdot,\cdot)$
- Latent posteriors are searched in the form:

$$\mathbb{Q}(\mathbf{f}\,|\,\mathcal{D}) = \mathcal{N}\left(\mathbf{f} \,\middle|\, \mathbf{m} + \mathbf{K}\alpha, (\mathbf{K}^{-1} + \mathbf{W})^{-1}\right) \quad (1)$$

- In the Gaussian likelihood case: $y_i \sim \mathcal{N}(f_i, \sigma_n^2)$; the inference is exact:

$$\mathbf{W} = \mathbf{I}\sigma_n^{-2}$$
$$\alpha = (\mathbf{K} + \mathbf{W}^{-1})^{-1}(\mathbf{y} - \mathbf{m}) = \text{solve}_\mathbf{K}(\mathbf{W}, \mathbf{r})$$
$$\log Z_{\text{GPR}} = -\frac{1}{2}\left[\alpha^\top \mathbf{r} + \text{ld}_\mathbf{K}(\mathbf{W}) + N \log(2\pi\sigma_n^2)\right]$$

- Direct application of these expression leads to $\mathcal{O}(n^3)$ computational complexity

Stochastic Differential Equation (SDE) Formulation

- Instead of working with the covariance function, the latent can be expressed in terms of an SDE
- This is a continuous-time model [3], [6]:

$$\frac{d\,\mathbf{f}(t)}{d\,t} = \mathbf{F}\,\mathbf{f}(t) + \mathbf{L}\,\mathbf{w}(t), \quad \mathbf{f}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{P}_\infty)$$

- $\mathbf{w}(t)$ is the driving multidimesional white noise
- The original latent can be evaluated at t by $f(t) = \mathbf{H}\,\mathbf{f}(t)$
- $\mathbf{F}, \mathbf{L}, \mathbf{H}, \mathbf{P}_\infty$ are determined from the covariance function

Discrete-Time State Space Model

- Solve the SDE between data points (equivalent discrete-time model):

$$\mathbf{f}_i = \mathbf{A}_{i-1}\,\mathbf{f}_{i-1} + \mathbf{q}_{i-1}; \quad \mathbf{q}_{i-1} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_{i-1})$$

- Parameters of discrete model:

$$\mathbf{A}_i = \mathbf{A}[\Delta t_i] = e^{\Delta t_i \mathbf{F}}, \qquad (2)$$
$$\mathbf{Q}_i = \mathbf{P}_\infty - \mathbf{A}_i\,\mathbf{P}_\infty\,\mathbf{A}_i^\top$$

- Advantages of the state space model:
  - Inference can be done in $\mathcal{O}(n)$ space and time complexity
  - This is done by running Kalman filtering (KF) and Rauch–Tung–Striebel (RTS) smoother algorithms
  - Evidence computation and its derivatives scales also as $\mathcal{O}(n)$

Fast Computation of $\mathbf{A}_i$ and $\mathbf{Q}_i$

Problem:

- Parameters of the solved SDE (2) depend on matrix exponents: $e^{\Delta t_i \mathbf{F}}$
- Many different Δt_i lead to expensive computation of matrix exponents

Solution:

- Mapping $\psi : s \mapsto e^{s\mathbf{X}}$ is smooth, hence use interpolation ideas (similar to KISS-GP [4])
- Evaluate $\psi : s \mapsto e^{s\mathbf{X}}$ on an equispaced grid $s_1, s_2, ..., s_K$, where $s_j = s_0 + j \cdot \Delta s$
- Use 4-point interpolation: $\mathbf{A} \approx c_1\mathbf{A}_{j-1} + c_2\mathbf{A}_j + c_3\mathbf{A}_{j+1} + c_4\mathbf{A}_{j+2}$. Coefficients $\{c_i\}_{i=1}^4$ are efficiently computable
- As shown below, this gives additional speed-up compared to the standard state space approch

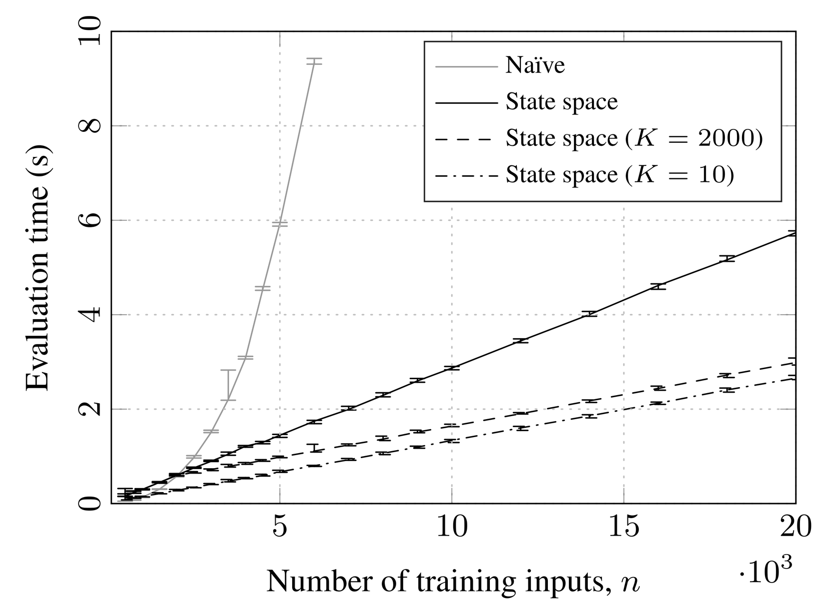

Figure 1: Empirical computational times of GP prediction using the GPML toolbox implementation as a function of number of training inputs, n, and degree of approximation, K.

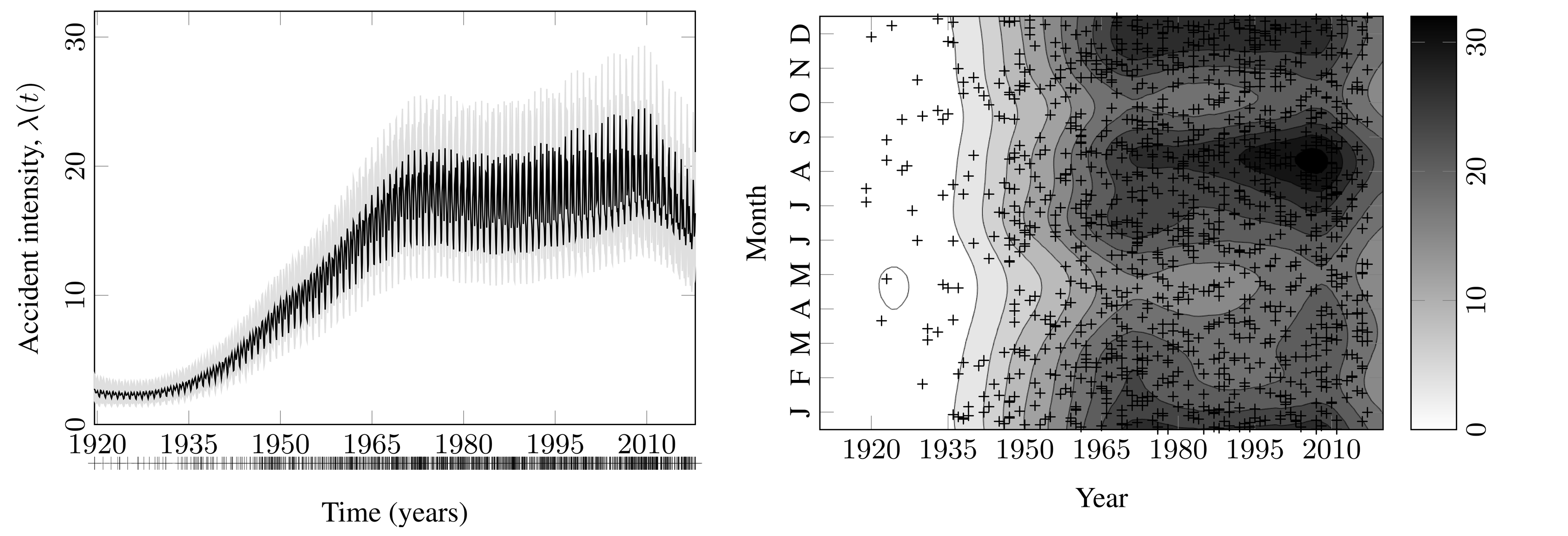

Figure 2: **(a)** Intensity of aircraft incidents modeled by a log Gaussian Cox process with the mean and approximate 90% confidence regions visualized ($N = 35{,}959$). **(b)** The time course of the seasonal effect in the airline accident intensity, plotted in a year vs. month plot (with wrap-around continuity between edges).

Computational Primitives

The following computational primitives allow to cast the covariance approximation in more generic terms:

- Linear system solving: $\text{solve}_\mathbf{K}(\mathbf{W}, \mathbf{r}) := (\mathbf{K} + \mathbf{W}^{-1})^{-1}\mathbf{r}$
- Matrix-vector multiplications: $\text{mvm}_\mathbf{K}(\mathbf{r}) := \mathbf{Kr}$
- Log-determinants: $\text{ld}_\mathbf{K}(\mathbf{W}) := \log|\mathbf{B}|$ with well-conditioned $\mathbf{B} = \mathbf{I} + \mathbf{W}^{\frac{1}{2}}\,\mathbf{K}\,\mathbf{W}^{\frac{1}{2}}$
- Predictions need latent mean $\mathbb{E}[f_*]$ and variance $\mathbb{V}[f_*]$

SpInGP:

- The first two computational primitives are calculated using *SpInGP* [5] approach:

$$\text{solve}_\mathbf{K}(\mathbf{W}, \mathbf{r}) = \mathbf{Wr} - \mathbf{WGR}^{-1}\mathbf{G}^\top\mathbf{Wr}$$
$$\text{mvm}_\mathbf{K}(\mathbf{r}) = \mathbf{GT}^{-1}\mathbf{QT}^{-\top}\mathbf{G}^\top\mathbf{r}$$

- Matrices $\mathbf{R}$, $\mathbf{T}$ and $\mathbf{Q}$ are defined via $\mathbf{A}_i$ and $\mathbf{Q}_i$, see paper [1]

KF and RTS Smoothing:

- The last two computational primitives are solved by Kalman filtering and RTS smoothing
- Predictions are computed by primitive 4 and then by propagation through likelihood

Comments:

- Derivatives of computational primites, required for learning, are computed in a similar way
- *SpInGP* involves computations with block-tridiagonal matrices. These computations are similar to KF and RTS smoothing (see paper [1] Appendix)

Approximate Inference

Laplace Approximation (LA):

- Second-order Taylor expansion around the mode of the posterior (1)
- Mode is found by Newton method
- Evidence approximation: $\log Z_{LA} = -\frac{1}{2}\left[\alpha^\top \text{mvm}_\mathbf{K}(\alpha) + \text{ld}_\mathbf{K}(\mathbf{W}) - 2\sum_i \log \mathbb{P}(y_i|\hat{f}_i)\right]$

Variational Bayes (VB):

- Lower bound the likelihood terms: $\log \mathbb{P}(y_i|f_i) = \max_{W_{ii}} b_i f_i - W_{ii} f_i^2/s + h(W_{ii})$
- Inference is cast as a sequence of LA with smoothed log likelihood
- $\log Z \geq \log Z_{\text{VB}} = -\frac{1}{2}\left[\alpha^\top \text{mvm}_\mathbf{K}(\alpha) + \text{ld}_\mathbf{K}(\mathbf{W}) - 2\sum_i \ell_{\text{VB}}(f_i) - 2\rho_{\text{VB}}\right]$

Direct KL minimization (KL):

- Gaussian approximation to the posterior is assumed. Variation lower bound is minimized
- Inference is cast as a sequence of GP regressions with convolved likelihood
- $\log Z \geq \log Z_{\text{KL}} = -\frac{1}{2}\left[\alpha^\top \text{mvm}_\mathbf{K}(\alpha) + \text{ld}_\mathbf{K}(\mathbf{W}) - 2\sum_i \ell_{\text{KL}}(f_i) - 2\rho_{\text{KL}}\right]$

Assumed Density Filtering (ADF):

- Equivalent to single-sweep Expectation Propagation (EP)
- Estimation of evidence and its derivatives in only one pass of Kalman filter

Experiments

- We show superior computational scaling with *exact* handling of the latent (Figure 1)
- A robust regression (Student's t likelihood) study example with $n = 34{,}154$ observations
- A new interesting data set with commercial airline accidents dates scraped from Wikipedia [7]
- Accidents over the time-span of $\sim$100 years, $n = 35{,}959$ days
- We model the accident intensity as a Log Gaussian Cox process (Poisson likelihood)
- The GP prior is set up as:

$$k(t,t') = k_{\text{Mat.}}(t,t') + k_{\text{per.}}(t,t')\,k_{\text{Mat.}}(t,t')$$

- Figure 2 shows the results for modelling the intensity of aircraft incidents

Discussion

- This paper brings together research done in state space GPs and non-Gaussian approximate inference
- We improve stability and provide additional speed-up by fast computations of the state space model transitions
- We provide unifying code for all approches in GPML toolbox v. 4.2

References

[1] H. Nickisch, A. Solin, and A. Grigorievskiy (2018). State Space Gaussian Processes with Non-Gaussian Likelihood. In *ICML*.

[2] C.E. Rasmussen and C.K.I. Williams (2006). *Gaussian Processes for Machine Learning.* The MIT Press.

[3] J. Hartikainen, and S. Särkkä (2010). Kalman filtering and smoothing solutions to temporal Gaussian process regression models. In *MLSP*.

[4] A. G. Wilson, and H. Nickisch (2015). Kernel Interpolation for Scalable Structured Gaussian Processes (KISS-GP). In *ICML*.

[5] A. Grigorievskiy and N. Lawrence and S. Särkkä (2017). Parallelizable Sparse Inverse Formulation Gaussian Processes (SpInGP). In *MLSP*.

[6] S. Särkkä and A. Solin (2018). Applied Stochastic Differential Equations. Cambridge University Press, Cambridge.

[7] Wikipedia (2018). URL `https://en.wikipedia.org/wiki/List_of_accidents_and_incidents_involving_commercial_aircraft`.

Codes

The code is published as part of the GPML toolbox:
`http://www.gaussianprocess.org/gpml/code/`

Contact details: hannes@nickisch.org and arno.solin@aalto.fi, ICML 2018, Stockholm, Sweden.