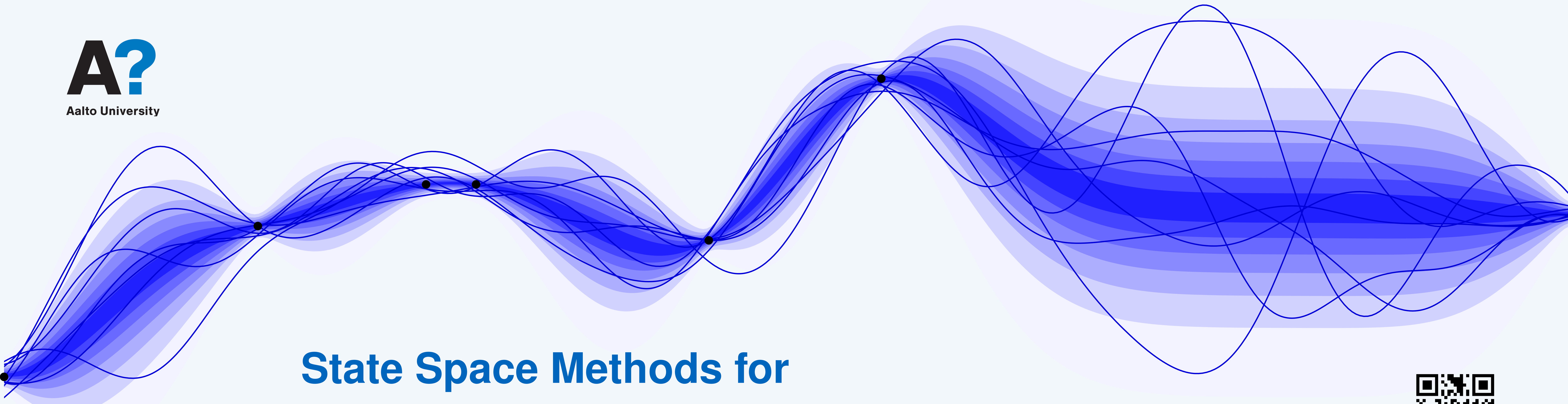


State Space Methods for Efficient Inference in Student-t Process Regression

Arno Solin Simo Särkkä

Aalto University, Finland

Introduction

- The flexibility of Student-t processes (TPs) over Gaussian processes (GPs) robustifies inference in noisy data [1,2].
- Predictive covariances explicitly depend on the training observations.
- For an entangled noise model, the canonical-form TP regression problem can be solved analytically [2].
- The naive TP and GP solutions share the same cubic computational cost in the number of training observations.
- We show how a large class of temporal TP regression models can be reformulated as state space models.
- We derive a forward filtering and backward smoothing recursion for doing the inference analytically in linear time complexity.

Student-t Processes

- In TP regression [2], we predict the output $f(t_*)$ with a known input $t_* \in \mathbb{R}$, given $\mathcal{D}_n = \{(t_k, y_k) \mid k = 1, 2, \ldots, n\}$:
$$f(t) \sim \mathcal{TP}(0, k(t,t'), \nu),$$
$$y_k = f(t_k).$$
- The direct solution to the TP regression problem gives predictions for the latent function
$$\mathbb{E}[f(t_*)] = \mathbf{k}_*^\mathsf{T} \mathbf{K}^{-1} \mathbf{y},$$
$$\mathbb{V}[f(t_*)] = \frac{\nu - 2 + \mathbf{y}^\mathsf{T}\mathbf{K}^{-1}\mathbf{y}}{\nu - 2 + n} \left(k_\theta(t_*, t_*) - \mathbf{k}_*^\mathsf{T}\mathbf{K}^{-1}\mathbf{k}_*\right).$$
- The noise model is included in the covariance function: $\mathbf{K}_{ij} = k_\theta(t_i, t_j) + \sigma_n^2 \delta_{i,j}$.
- The computational scaling is $\mathcal{O}(n^3)$ due to the matrix inverse.
- We call this the 'naive' way of solving the inference problem and derive an alternative approach in what follows.

State Space Model

- Stationary Gaussian processes with a rational spectra can be converted to in law equivalent state space stochastic differential equations (SDEs) [3].
- These state space SDEs can be written as
$$\frac{\mathrm{d}\mathbf{f}(t)}{\mathrm{d}t} = \mathbf{F}\mathbf{f}(t) + \mathbf{L}\mathbf{w}(t), \quad \text{and} \quad f(t_k) = \mathbf{H}\mathbf{f}(t_k),$$
where $\mathbf{f}(t) = (f_1(t), f_2(t), \ldots, f_m(t))^\mathsf{T}$ holds the m stochastic processes, and $\mathbf{w}(t)$ is a white noise process with spectral density $\mathbf{Q}_\mathrm{c}$, and initial state $\mathbf{f}(0) \sim \mathrm{N}(\mathbf{0}, \mathbf{P}_0)$.
- A TP can be constructed as a scale mixture of state space form SDEs by setting the spectral density to $\gamma\mathbf{Q}_\mathrm{c}$, and using the initial state $\mathbf{f}(0) \sim \mathrm{N}(\mathbf{0}, \gamma\mathbf{P}_0)$, where γ is an inverse gamma random variable.
- The solution can be written out in closed-form at the specified time points t_k, $k = 1, 2, \ldots$, as $\mathbf{f}(t_k) = \mathbf{f}_k$ such that $\mathbf{f}_0 \sim \mathrm{N}(\mathbf{0}, \gamma\mathbf{P}_0)$ and
$$\mathbf{f}_k = \mathbf{A}_{k-1}\mathbf{f}_{k-1} + \mathbf{q}_{k-1},$$
where $\mathbf{q}_{k-1} \sim \mathrm{N}(\mathbf{0}, \gamma\mathbf{Q}_{k-1})$.
- The entangled noise model is included by augmenting it into the state.

Student-t Filtering and Smoothing

- Filtering and smoothing [4] in state space models refer to the Bayesian methodology of computing posterior distributions of the latent state based on a history of noisy measurements.
- Filtering distributions are the marginal distributions of the state $\mathbf{f}_k$ given the current and previous measurements up to the point t_k: $\mathbf{f}_k \mid \mathcal{D}_k \sim \mathrm{MVT}(\mathbf{m}_{k|k}, \mathbf{P}_{k|k}, \nu_k)$ (see Alg. 1).
- Prediction distributions are the marginal distributions of the future state following the previous observation: $\mathbf{f}_{k+j} \mid \mathcal{D}_k \sim \mathrm{MVT}(\mathbf{m}_{k+j|k}, \mathbf{P}_{k+j|k}, \nu_k)$ (see Alg. 1).
- Smoothing distributions are the marginal distributions of the state given all the measurements in the interval: $\mathbf{f}_k \mid \mathcal{D}_n \sim \mathrm{MVT}(\mathbf{m}_{k|n}, \mathbf{P}_{k|n}, \nu_n)$ (see Alg. 2).
- The filter gives the marginal likelihood for hyperparameter optimization.
- The smoothing outcome corresponds to the naive TP regression result.

Algorithm 1: Student-t filter.

for $k = 1, 2 \ldots, n$ **do**

Filter prediction:
$$\mathbf{m}_{k|k-1} = \mathbf{A}_{k-1}\mathbf{m}_{k-1|k-1}$$
$$\mathbf{P}_{k|k-1} = \mathbf{A}_{k-1}\mathbf{P}_{k-1|k-1}\mathbf{A}_{k-1}^\mathsf{T} + \gamma_{k-1}\mathbf{Q}_{k-1}$$
Filter update:
$$\mathbf{v}_k = \mathbf{y}_k - \mathbf{H}_k\mathbf{m}_{k|k-1}$$
$$\mathbf{S}_k = \mathbf{H}_k\mathbf{P}_{k|k-1}\mathbf{H}_k^\mathsf{T}$$
$$\gamma_k = \frac{\gamma_{k-1}}{\nu_k - 2}(\nu_{k-1} - 2 + \mathbf{v}_k^\mathsf{T}\mathbf{S}_k^{-1}\mathbf{v}_k)$$
$$\mathbf{K}_k = \mathbf{P}_{k|k-1}\mathbf{H}_k^\mathsf{T}\mathbf{S}_k^{-1}$$
$$\mathbf{m}_{k|k} = \mathbf{m}_{k|k-1} + \mathbf{K}_k\mathbf{v}_k$$
$$\mathbf{P}_{k|k} = \frac{\gamma_k}{\gamma_{k-1}}(\mathbf{P}_{k|k-1} - \mathbf{K}_k\mathbf{S}_k\mathbf{K}_k^\mathsf{T})$$
end

Algorithm 2: Student-t smoother.

for $k = n-1, n-2, \ldots, 1$ **do**

Smoother prediction:
$$\mathbf{m}_{k+1|k} = \mathbf{A}_k\mathbf{m}_{k|k}$$
$$\mathbf{P}_{k+1|k} = \mathbf{A}_k\mathbf{P}_{k|k}\mathbf{A}_k^\mathsf{T} + \gamma_k\mathbf{Q}_k$$
Smoother update:
$$\mathbf{G}_k = \mathbf{P}_{k|k}\mathbf{A}_k^\mathsf{T}\mathbf{P}_{k+1|k}^{-1}$$
$$\mathbf{m}_{k|n} = \mathbf{m}_{k|k} + \mathbf{G}_k(\mathbf{m}_{k+1|n} - \mathbf{m}_{k+1|k})$$
$$\mathbf{P}_{k|n} = \frac{\gamma_n}{\gamma_k}(\mathbf{P}_{k|k} - \mathbf{G}_k\mathbf{P}_{k+1|k}\mathbf{G}_k^\mathsf{T}) + \mathbf{G}_k\mathbf{P}_{k+1|n}\mathbf{G}_k^\mathsf{T}$$
end

Conclusions

- We have generalized the connection between Gaussian process regression and Kalman filtering to more general elliptical processes and non-Gaussian Bayesian filtering.
- This link enables the use of efficient sequential inference methods to solve TP regression problems in $\mathcal{O}(n)$ time complexity.
- An example implementation is available on the author web page:

http://arno.solin.fi

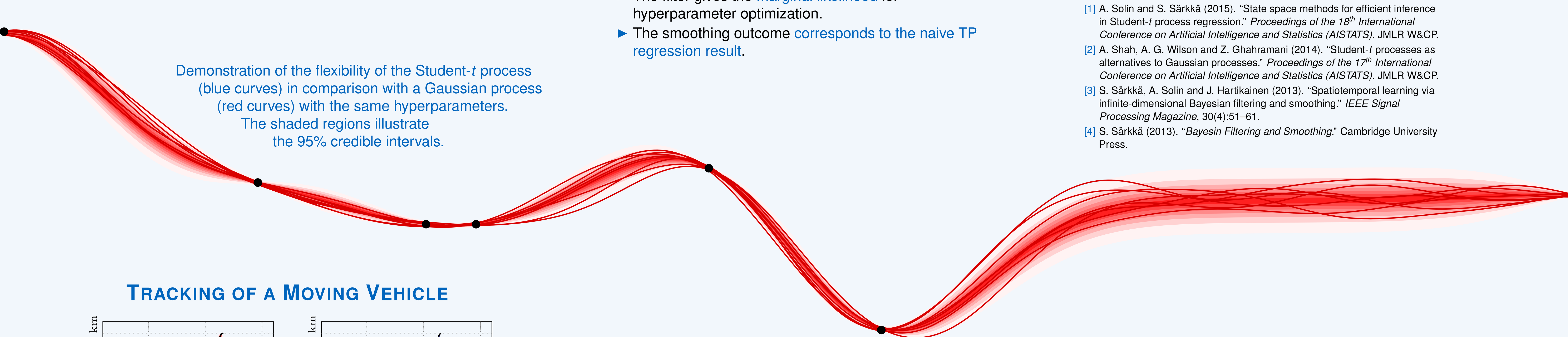
References

[1] A. Solin and S. Särkkä (2015). "State space methods for efficient inference in Student-t process regression." *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*. JMLR W&CP.

[2] A. Shah, A. G. Wilson and Z. Ghahramani (2014). "Student-t processes as alternatives to Gaussian processes." *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*. JMLR W&CP.

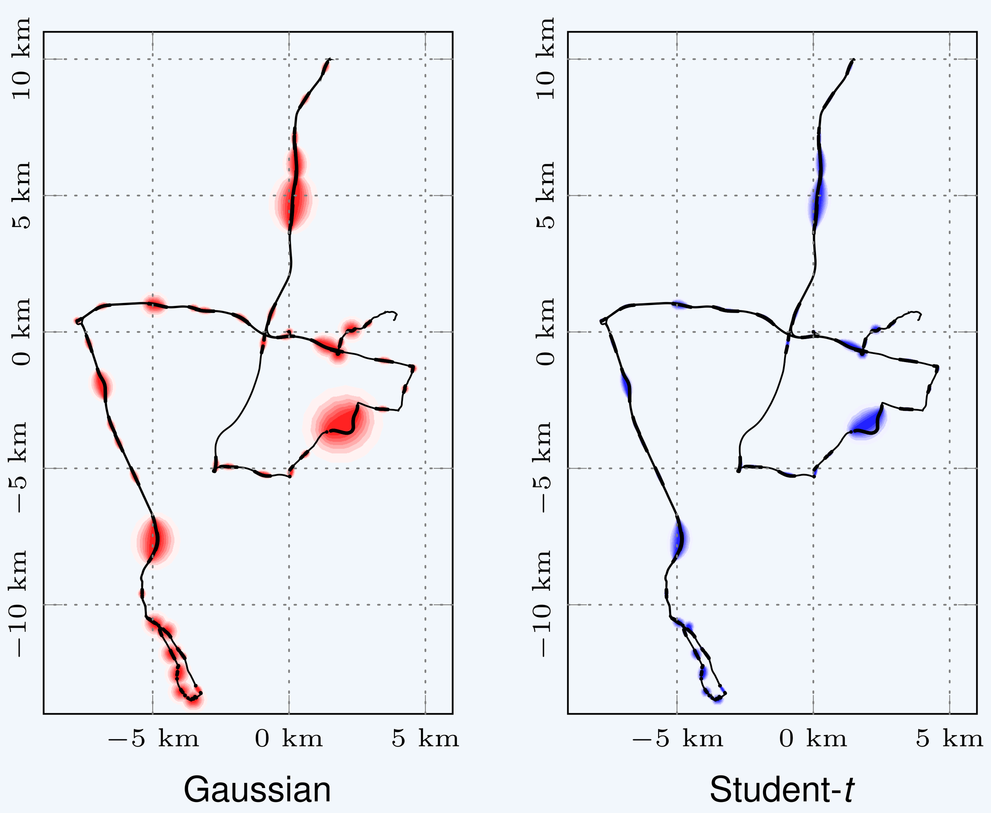
[3] S. Särkkä, A. Solin and J. Hartikainen (2013). "Spatiotemporal learning via infinite-dimensional Bayesian filtering and smoothing." *IEEE Signal Processing Magazine*, 30(4):51–61.

[4] S. Särkkä (2013). "*Bayesin Filtering and Smoothing*." Cambridge University Press.



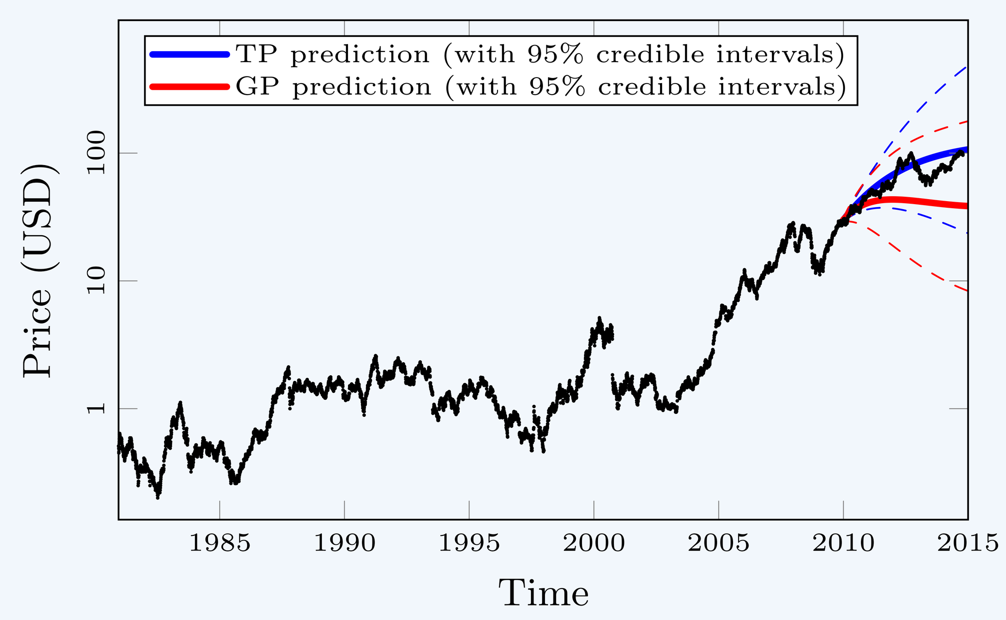
Demonstration of the flexibility of the Student-t process (blue curves) in comparison with a Gaussian process (red curves) with the same hyperparameters. The shaded regions illustrate the 95% credible intervals.

Tracking of a Moving Vehicle



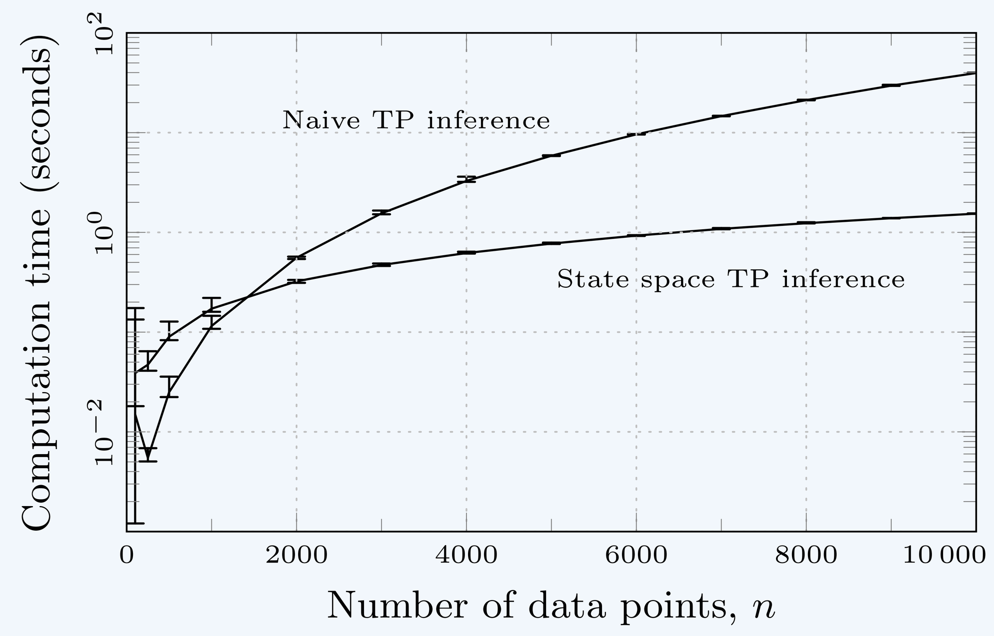
Interpolation of missing GPS observations by two-dimensional GP regression (Gaussian smoothing) and TP regression (Student-t smoothing). The unknown ground truth is shown by dots and the colored patches illustrate the credible intervals up to 95%.

Stock Price Data



The log share price of Apple Inc. ($n = 8\,537$) modeled by GP/TP with a covariance function sum of a constant, linear, Matérn (smoothness 3/2), and exponential covariance function. The main difference comes from the different hyperparameters.

Computational Efficiency



Demonstration of the computational benefits of the state space model in solving a TP regression problem for a number of data points up to 10 000.