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

Arno Solin
arno.solin@aalto.fi
Aalto University

Simo Särkkä
simo.sarkka@aalto.fi
Aalto University

Abstract

The added flexibility of Student-t processes (TPs) over Gaussian processes (GPs) robustifies inference in outlier-contaminated noisy data. The uncertainties are better accounted for than in GP regression, because the predictive covariances explicitly depend on the training observations. For an entangled noise model, the canonical-form TP regression problem can be solved analytically, but 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, and how a forward filtering and backward smoothing recursion can be derived for solving the inference analytically in linear time complexity. This is a novel finding that generalizes the previously known connection between Gaussian process regression and Kalman filtering to more general elliptical processes and non-Gaussian Bayesian filtering. We derive this connection, demonstrate the benefits of the approach with examples, and finally apply the method to empirical data.

1 INTRODUCTION

Gaussian processes (GPs, [1]) provide a flexible way of imposing non-parametric priors over functions, which has made them popular modeling tools in both Bayesian machine learning and signal processing. In signal processing, temporal GPs are typically represented as state space models [2], whereas the kernel (covariance function) formalism is favored in machine learning. The link between these two representations is interesting, because it enables the combination of the intuitive model specification from machine learning with computationally efficient signal processing methods [3–5]. Most notably, for one-dimensional models this reduces the computational cost of a naive GP regression solution from $\mathcal{O}(n^3)$ to $\mathcal{O}(n)$ in the number of training data points n by solving the inference problem by Kalman filtering methods (see, *e.g.*, [6]).

The success of Gaussian processes has awakened an interest in expanding the methodology to more general families of elliptical processes [7, 8], such as the Student-t process (TP, see, *e.g.*, [9, 10]). The scale mixture connection to Gaussian processes justifies the added flexibility in TPs, and longer tails provide robustness against outliers. Additionally, the predictive covariance explicitly depends on the training observations, even though in GPs it only depends on the training inputs. Hence, noise in the measurements also gives information on the uncertainty, whereas in GP regression this information is ignored. Recently Shah *et al.* [10] resurrected the concept of TP regression, where they proposed the noise model to be included in the kernel function. They showed that contrary to the TP described by Rasmussen and Williams [1], this model can be solved analytically and it was shown to be beneficial in comparison to the GP regression approach.

In this paper, we present a connection between temporal Student-t processes and state space models, and propose a novel method for solving the TP regression problem recursively in linear time complexity with respect to the number of observations. This generalizes the known Kalman filtering connection of GPs [3–5] to more general elliptical processes. As the degrees of freedom $\nu \to \infty$, we recover the Kalman filter and the model reverts to GP regression. We construct the Student-t processes analogously to Shah *et al.* [10] as a mixture of Gaussian processes with only a single in-

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

verse Wishart random variable as the scaling distribution. By representing the Student-t distributions as a scale mixture, marginalizations in the Kalman smoother forward–backward algorithm can be interchanged with marginalization over the scale variable. It is also possible to construct Student-t processes (*i.e.*, processes with Student-t marginals) by other means (see [8]): by constructing a suitable Lévy subordinator which transforms a Gaussian process into a Student-t process via a time change, or by constructing a stochastic differential equation with a suitable non-linear drift. Our approach roughly corresponds to using a globally scaled time change instead of stochastic subordinator-based time scaling. Even though this hints that a generalization to Lévy subordinator based processes might be possible, it is non-trivial due to the entanglement of the process and noise.

The problem of robustifying GP regression models has been tackled before, mostly by considering long-tailed measurement noise. In signal processing several types of robust filtering approaches exist (see, *e.g.*, [11–15], and references therein), whereas under the GP framework in machine learning, robustification has been implemented by a Student-t observation model [16]. Similar Student-t likelihood based methods in state space models have been analyzed in [12, 13]. However, in these approaches the inference is approximate, whereas this paper considers analytic solutions enabled by the noise entanglement.

The main contributions of this paper are:

- A novel explicit connection between Student-t processes and state space models.
- A novel Bayesian filtering and smoothing based inference scheme for solving temporal Student-t process regression problems in linear time complexity.

This paper is structured as follows. In Section 2, we begin by introducing the Student-t process and define the required concepts for constructing the inference scheme. In Section 3 we establish the connection between the Student-t process and the corresponding state space model and describe our inference algorithm. Numerical experiments are conducted in Section 4, and the results are discussed in Section 5.

2 STUDENT-t PROCESSES

In this section, we provide the required backdrop for the rest of the paper by considering the properties of the Student-t distribution and processes. A survey on concepts and results related to the Student-t distribution has been provided by Kotz and Nadarajah [17], and here we only briefly present the relevant aspects. For consistency between the multivariate Gaussian (*i.e.*, the GP parametrization) and the multivariate Student-t parametrization (*i.e.*, the TP parametrization), we define the multivariate Student-t distribution as follows.

Definition 2.1. *The random variable $\mathbf{y} \in \mathbb{R}^n$ is multivariate Student-t distributed, $\mathbf{y} \sim \mathrm{MVT}(\boldsymbol{\mu}, \mathbf{K}, \nu)$, with degrees of freedom $\nu > 2$, mean $\boldsymbol{\mu} \in \mathbb{R}^n$, and covariance matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$, if it has the density function*

$$\mathrm{MVT}(\mathbf{y} \mid \boldsymbol{\mu}, \mathbf{K}, \nu) = \frac{\Gamma(\frac{\nu+n}{2})}{\Gamma(\frac{\nu}{2})} \frac{1}{((\nu-2)\pi)^{\frac{n}{2}}} \frac{1}{|\mathbf{K}|^{\frac{1}{2}}} \left(1 + \frac{1}{\nu-2}(\mathbf{y}-\boldsymbol{\mu})^{\mathsf{T}} \mathbf{K}^{-1} (\mathbf{y}-\boldsymbol{\mu})\right)^{-\frac{\nu+n}{2}}. \quad (1)$$

From Definition 2.1, as $\nu \to \infty$, we recover the multivariate Gaussian density, $\mathrm{N}(\mathbf{y} \mid \boldsymbol{\mu}, \mathbf{K})$. The Student-$t$ distribution can be defined as a scale mixture of Gaussians (see, *e.g.*, [1, 8, 9]), as in the following lemma.

Lemma 2.2. *Let $\gamma \sim \mathrm{IG}(\nu/2, (\nu-2)/2)$ be inverse gamma distributed and $\mathbf{y} \mid \gamma \sim \mathrm{N}(\boldsymbol{\mu}, \gamma\mathbf{K})$ be a Gaussian with mean $\boldsymbol{\mu}$ and scaled covariance $\gamma\mathbf{K}$, then marginally $\mathbf{y} \sim \mathrm{MVT}(\boldsymbol{\mu}, \mathbf{K}, \nu)$.*

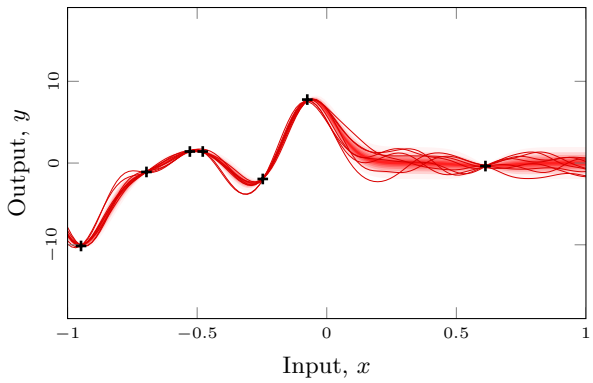
A sketch for a proof for the above lemma is provided in the supplementary material of this paper. The same result can be derived by placing an inverse Wishart process prior on the kernel function, leading to a Student-t process [10]. The Student-t distribution inherits several appealing features from the Gaussian, the most important in this context being an analytic conditional distribution. The following result can be readily found in literature (see, *e.g.*, [17]):

Lemma 2.3. *Let $\mathbf{f}_1 \in \mathbb{R}^{n_1}$ and $\mathbf{f}_2 \in \mathbb{R}^{n_2}$ be jointly t distributed with ν degrees of freedom. The conditional density for a multivariate Student-t has an analytic form: $\mathbf{f}_1 \mid \mathbf{f}_2 \sim \mathrm{MVT}\big(\boldsymbol{\mu}_{1|2}, \mathbf{K}_{1|2}, \nu_{1|2}\big)$, with mean $\boldsymbol{\mu}_{1|2} = \mathbf{K}_{12}\mathbf{K}_{22}^{-1}(\mathbf{f}_2 - \boldsymbol{\mu}_2) + \boldsymbol{\mu}_1$, covariance $\mathbf{K}_{1|2} = \frac{\nu-2+\beta}{\nu-2+n_2}\big(\mathbf{K}_{11} - \mathbf{K}_{12}\mathbf{K}_{22}^{-1}\mathbf{K}_{21}\big)$, $\beta = (\mathbf{f}_2-\boldsymbol{\mu}_2)^{\mathsf{T}}\mathbf{K}_{22}^{-1}(\mathbf{f}_2 - \boldsymbol{\mu}_2)$, and degrees of freedom $\nu_{1|2} = \nu + n_2$.*

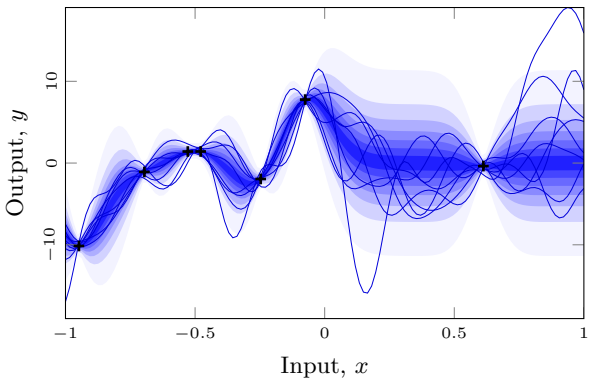
We define the Student-t process using a similar parametrization as Shah *et al.* [10]:

Definition 2.4. *The process $f(\mathbf{x})$ is a Student-t process, $f(\mathbf{x}) \sim \mathcal{TP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'), \nu)$, on $\mathcal{X}$ with degrees of freedom $\nu > 2$, a mean function $\mu : \mathcal{X} \to \mathbb{R}$, and a covariance function (kernel) $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, if any finite collection of function values has a joint multivariate Student-t distribution such that $(f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_n))^{\mathsf{T}} \sim \mathrm{MVT}(\boldsymbol{\mu}, \mathbf{K}, \nu)$, where $\mathbf{K}_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$ and $\boldsymbol{\mu}_i = \mu(\mathbf{x}_i)$, for $i, j = 1, 2, \ldots, n$.*

The Student-t process generalizes the concept of Gaussian processes. For a Student-t process $f \sim$



(a) Samples from the GP posterior



(b) Samples from the TP posterior

Figure 1: Demonstration of the added flexibility of the Student-t process in comparison with a Gaussian process with the same hyperparameter values for an exponentiated quadratic (squared exponential) covariance function. The shaded regions illustrate the 95% credible intervals.

$\mathcal{TP}(\mu(\cdot), k(\cdot,\cdot), \nu)$, as $\nu \to \infty$, we recover the Gaussian process with the corresponding mean and covariance functions. Shah *et al.* argue that the TP is the most general elliptical process which has an analytically representable density. Figure 1 illustrates the difference between a GP and TP by presenting random draws from the posterior distributions. The exponentiated quadratic (squared exponential, Gaussian, RBF) covariance function was given the same hyperparameter values in both models.

2.1 TP Regression as a Generalization of GP Regression

We consider the concept of TP regression [10], which is concerned with predicting an unknown scalar output $f(\mathbf{x}_*)$ associated with a known input $\mathbf{x}_* \in \mathbb{R}^d$, given a training data set $\mathcal{D} = \{(\mathbf{x}_k, y_k) \mid k = 1, 2, \ldots, n\}$. The model function is assumed to be a realization of a zero-mean Student-t random process prior (with covariance function $k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}')$) and the observations corrupted by an entangled Student-t noise process:

$$\begin{aligned} f(\mathbf{x}) &\sim \mathcal{TP}(0, k(\mathbf{x}, \mathbf{x}'), \nu), \\ y_k &= f(\mathbf{x}_k), \end{aligned} \tag{2}$$

where the noise model is incorporated in the covariance function. The direct solution to the TP regression problem gives predictions for the latent function, $p(f(\mathbf{x}_*) \mid \mathbf{x}_*, \mathcal{D}) = \mathrm{MVT}(\mathbb{E}[f(\mathbf{x}_*)], \mathbb{V}[f(\mathbf{x}_*)], \nu + n)$. By Lemma 2.3, this can be computed in closed-form as

$$\mathbb{E}[f(\mathbf{x}_*)] = \mathbf{k}_*^{\mathsf{T}} \mathbf{K}^{-1} \mathbf{y}, \tag{3}$$

$$\mathbb{V}[f(\mathbf{x}_*)] = \frac{\nu - 2 + \mathbf{y}^{\mathsf{T}} \mathbf{K}^{-1} \mathbf{y}}{\nu - 2 + n} \left(k_{\boldsymbol{\theta}}(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^{\mathsf{T}} \mathbf{K}^{-1} \mathbf{k}_* \right),$$

where $\mathbf{k}_*$ is an n-dimensional vector with the ith entry being $k_{\boldsymbol{\theta}}(\mathbf{x}_*, \mathbf{x}_i)$, and $\mathbf{y}$ is a vector of the n observations. The noise model is included in the covariance function by adding a noise covariance function to the parametrized kernel: $\mathbf{K}_{ij} = k_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_j) + \sigma_{\mathrm{n}}^2 \delta_{i,j}$, where $\delta_{i,j}$ is the Kronecker delta. As pointed out in [10], the noise will be uncorrelated with the latent function, but not independent. In the limit $\nu \to \infty$, this model tends to a GP regression model with independent Gaussian noise. The computational complexity of these equations is inherently cubic in n due to the term $\mathbf{K}^{-1}$.

Training the model amounts to estimating the hyperparameters $\boldsymbol{\theta}$ of the covariance function $k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}')$, the noise scale σ_{n}^2, and degrees of freedom ν. In practice, this is often done by minimizing the negative log marginal likelihood function (following from Def. 2.1):

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= -\log p(\mathbf{y} \mid \boldsymbol{\theta}, \nu) \\ &= \frac{n}{2} \log((\nu - 2)\pi) + \frac{1}{2} \log(|\mathbf{K}|) - \log \Gamma\left(\frac{\nu + n}{2} \right) \\ &\quad + \log \Gamma\left(\frac{\nu}{2} \right) + \frac{\nu + n}{2} \log\left(1 + \frac{\beta}{\nu - 2} \right), \end{aligned} \tag{4}$$

where $\beta = (\mathbf{y} - \boldsymbol{\mu})^{\mathsf{T}} \mathbf{K}^{-1} (\mathbf{y} - \boldsymbol{\mu})$. The power of the TP regression scheme over GP regression comes from the different marginal likelihood, leading to different hyperparameter values under training. Unfortunately, due to the appearance of $\mathbf{K}^{-1}$, evaluation of the marginal likelihood has a cubic computational complexity.

3 STUDENT-t PROCESSES AS STATE SPACE MODELS

In this section, we propose a novel way of reformulating the TP regression scheme from the previous section as a state space inference problem. We consider how the Student-t process can be written as a scaled mixture of linear stochastic differential equations which can be solved at a given finite number of input values.

We now restrict our interest to temporal processes, such that $\mathbf{x} = t$ ($\mathcal{X} = \mathbb{R}$). The following results could be generalized to spatio-temporal models similar to the generalization of spatio-temporal GPs (see [5]), where the model is seen as a Student-t field evolving in time.

3.1 Student-t Processes as Solutions to State Space Stochastic Differential Equations

Instead of directly working with the kernel formalism of the Student-t process $f(t)$ given in Definition 2.4, certain classes of covariance functions can be constructed as solutions of mth order linear stochastic differential equations (SDEs) or equivalently m-dimensional state space form SDEs.

Lemma 3.1. *Given a class of parametric stationary Gaussian processes with a rational spectrum, there exists a class of state space form SDEs which are weakly (in law) equivalent with this class of Gaussian processes, in the sense that their covariance functions match. These state space SDEs can be written as*

$$\frac{\mathrm{d}\mathbf{f}(t)}{\mathrm{d}t} = \mathbf{F}\mathbf{f}(t) + \mathbf{L}\mathbf{w}(t), \quad \textit{and} \quad f(t_k) = \mathbf{H}\mathbf{f}(t_k), \tag{5}$$

where $\mathbf{f}(t) = \big(f_1(t), f_2(t), \ldots, f_m(t)\big)^\mathsf{T}$ *holds the* m *stochastic processes, and* $\mathbf{w}(t)$ *is a multi-dimensional white noise process with spectral density* $\mathbf{Q}_\mathrm{c}$*, and initial state* $\mathbf{f}(0) \sim \mathrm{N}(\mathbf{0}, \mathbf{P}_0)$.

Proof. This follows from the constructions given in [5]. □

Theorem 3.2. *A Student-t process,* $f(t) \sim \mathcal{TP}(0, k(t,t'), \nu)$*, where* $k(t,t')$ *corresponds to a rational spectral density, can be constructed as a scale mixture of state space form SDEs in the form considered in Lemma 3.1 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.*

Proof. With the white noise spectral density $\gamma\mathbf{Q}_\mathrm{c}$, the spectral density of the process and hence its covariance function is directly proportional to γ while the mean remains zero. The result follows from Lemmas 2.2 and 3.1 together with the fact that the scaled mixtures will be weakly equivalent provided that the scale members are. □

As well-known, the continuous-time linear time-invariant model (5) can be solved for discrete points [2, 3]. In construction of the TP above, the model is defined by the feedback matrix $\mathbf{F}$, the noise effect matrix $\mathbf{L}$, the spectral density $\gamma\mathbf{Q}_\mathrm{c}$ of the white-noise process, the observation model $\mathbf{H}$, and the initial state covariance $\gamma\mathbf{P}_0$. The solution to (5) 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

$$\mathbf{f}_k = \mathbf{A}_{k-1}\mathbf{f}_{k-1} + \mathbf{q}_{k-1}, \tag{6}$$

where $\mathbf{q}_{k-1} \sim \mathrm{N}(\mathbf{0}, \gamma\mathbf{Q}_{k-1})$. The state transition and process noise covariance matrices can be solved analytically (see, *e.g.*, [5]):

$$\mathbf{A}_k = \boldsymbol{\Phi}(\Delta t_k) \quad \text{and} \tag{7}$$

$$\mathbf{Q}_k = \int_0^{\Delta t_k} \boldsymbol{\Phi}(\Delta t_k - \tau)\mathbf{L}\mathbf{Q}_\mathrm{c}\mathbf{L}^\mathsf{T}\boldsymbol{\Phi}(\Delta t_k - \tau)^\mathsf{T}\,\mathrm{d}\tau, \tag{8}$$

where $\Delta t_k = t_{k+1} - t_k$ and $\boldsymbol{\Phi}(\tau) = \exp(\mathbf{F}\tau)$ is the matrix exponential of the feedback matrix. For stationary models, the initial state covariance $\mathbf{P}_0$ is defined by the stationary covariance $\mathbf{P}_\infty$ that is the solution to the corresponding Lyapunov equation: $\frac{\mathrm{d}\mathbf{P}_\infty}{\mathrm{d}t} = \mathbf{F}\mathbf{P}_\infty + \mathbf{P}_\infty\mathbf{F}^\mathsf{T} + \mathbf{L}\mathbf{Q}_\mathrm{c}\mathbf{L}^\mathsf{T} = \mathbf{0}$. For these models, the state transition covariance matrix is given by $\mathbf{Q}_k = \mathbf{P}_\infty - \mathbf{A}_k\mathbf{P}_\infty\mathbf{A}_k^\mathsf{T}$.

Remark 3.3. *The results can be generalized to some classes of non-stationary models. See the results by Van Trees [18, Sec. A.3] and Anderson et al. [19].*

Remark 3.4. *Above, we have restricted the class of covariance functions to those with a rational spectrum, which excludes, for example, the exponentiated quadratic (squared exponential), rational quadratic, and quasi-periodic covariance functions. However, it has been recently shown in [20–22] that these covariance functions can be approximated to an arbitrary accuracy via finite-dimensional state space models.*

3.2 Including the Noise Model

The entangled measurement noise is augmented into the state as an additional component:

$$f_k = q_{k-1}, \quad \text{where} \quad q_{k-1} \sim \mathrm{N}(0, \sigma_\mathrm{n}^2). \tag{9}$$

This corresponds to the noise covariance function, $k_\mathrm{noise}(t_i, t_j) = \sigma_\mathrm{n}^2\,\delta_{i,j}$, where $\delta_{i,j}$ is the Kronecker delta function, and results in a diagonal Gram matrix.

This model can be derived by consider an Ornstein–Uhlenbeck process and its covariance function (also known as the exponential covariance function, [1]) $k_\mathrm{exp}(t,t') = \sigma_\mathrm{n}^2 \exp(-\lambda|t - t'|)$. The corresponding state space model can be given as a one-dimensional stochastic differential equation: $\frac{\mathrm{d}f(t)}{\mathrm{d}t} = -\lambda f(t) + w(t)$, where $w(t)$ is a white noise process with spectral density $Q_\mathrm{c} = 2\lambda\sigma_\mathrm{n}^2$, and the stationary state variance $P_\infty = \sigma_\mathrm{n}^2$. For this model, as $\lambda \to \infty$ (the characteristic length-scale going to zero), the model tends to a white noise process, which has the formal discrete-time solution given in Equation (9) (with zero feedback for any Δt).

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

As summing covariance functions, $k(t,t') = k_{\boldsymbol{\theta}}(t,t') + k_{\text{noise}}(t,t')$, under the kernel formalism corresponds to stacking state variables in the state space model, including the entangled noise contribution can be accomplished by augmenting the white noise process into the state variable. This leads to the following joint state space model: $\mathbf{F} = \text{blkdiag}(\mathbf{F}_{\boldsymbol{\theta}}, -\infty)$ and $\mathbf{P}_0 = \text{blkdiag}(\mathbf{P}_{\boldsymbol{\theta},0}, \sigma_{\text{n}}^2)$, and the observation model $\mathbf{H} = (\mathbf{H}_{\boldsymbol{\theta}}, 1)$ (training) and $\mathbf{H} = (\mathbf{H}_{\boldsymbol{\theta}}, 0)$ (prediction of the latent function).

3.3 Sequential Inference

Filtering and smoothing (see, *e.g.*, [6]) in state space models refer to the Bayesian methodology of computing posterior distributions of the latent state based on a history of noisy measurements. Let the observed data be denoted as $\mathcal{D}_n = \{(t_i, y_i) \mid i = 1, 2, \ldots, n\}$. In Bayesian filtering and smoothing the interest is put into the following marginal distributions:

- The *filtering distributions* are the outcome of the Bayesian filter. They are the marginal distributions of the state $\mathbf{f}_k$ given the current and previous measurements up to the point t_k: $\mathbf{f}_k \mid \mathcal{D}_k \sim \text{MVT}(\mathbf{m}_{k|k}, \mathbf{P}_{k|k}, \nu_k)$.

- The *prediction distributions*, which can be computed with the prediction step of the Bayesian filter, are the marginal distributions of the future state $\mathbf{f}_{k+j}$, for $j = 1, 2, \ldots$ steps following the previous observation: $\mathbf{f}_{k+j} \mid \mathcal{D}_k \sim \text{MVT}(\mathbf{m}_{k+j|k}, \mathbf{P}_{k+j|k}, \nu_k)$.

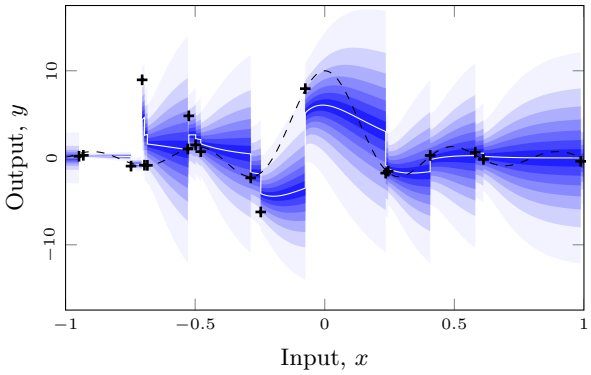
- The *smoothing distributions* computed by the Bayesian smoother are the marginal distributions of the state $\mathbf{f}_k, k = 1, 2, \ldots, n$ given all the measurements in the interval: $\mathbf{f}_k \mid \mathcal{D}_n \sim \text{MVT}(\mathbf{m}_{k|n}, \mathbf{P}_{k|n}, \nu_n)$. The smoothing solution corresponds to the naive solution in Equation (3).

Given the class of Student-t processes in Theorem 3.2, the TP regression problem in Equation (3) can be solved by sequentially solving a forward filtering problem, and updating the filtering outcome by running a backward smoother. This is constructed by sequentially predicting the next step as given by the link in Theorem 3.2, and updating the state as by Lemma 2.3.

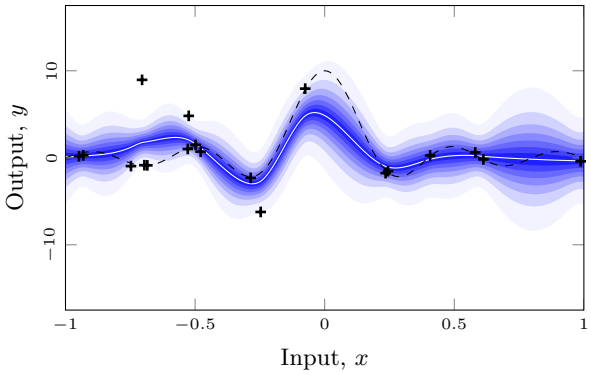
The inference scheme is presented as a closed-form recursion in Algorithm 1 (filter) and Algorithm 2 (smoother). The initial degrees of freedom are $\nu_0 = \nu$, scaling factor $\gamma_0 = 1$, prior state mean $\mathbf{m}_{0|0} = \mathbf{0}$, and prior state covariance $\mathbf{P}_{0|0} = \mathbf{P}_0$. The smoother is initialized by the filtering outcome. The degrees of freedom parameter is updated as $\nu_k = \nu_{k-1} + n_k$, where $n_k = 1$, if there is an update on time-step k, and $n_k = 0$ otherwise (for prediction of test points). Prediction of test inputs corresponds to including t_* in the filtering and smoothing sweeps, but skipping the filter update for the point. For training the hyperparameters, the negative log marginal likelihood can be evaluated sequentially as a by-product of the filtering recursion in Algorithm 1:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{k=1}^{n}\left\{\frac{1}{2}\log((\nu-2)\pi) + \frac{1}{2}\log(|\mathbf{S}_k|) + \log\Gamma\left(\frac{\nu_{k-1}}{2}\right) - \log\Gamma\left(\frac{\nu_k}{2}\right) + \frac{1}{2}\log\left(\frac{\nu_{k-1}-2}{\nu-2}\right) + \frac{\nu_k}{2}\log\left(1 + \frac{\mathbf{v}_k^{\mathsf{T}}\mathbf{S}_k^{-1}\mathbf{v}_k}{\nu_{k-1}-2}\right)\right\}, \tag{10}$$

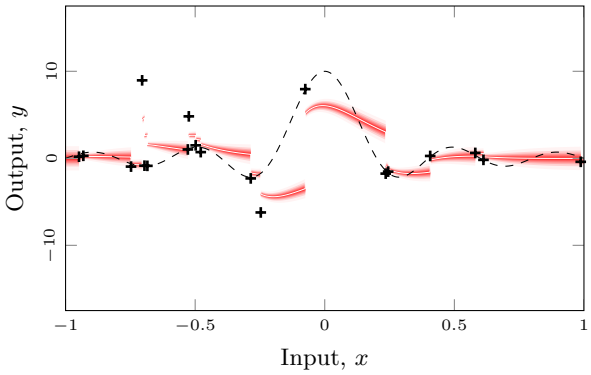
where $\mathbf{v}_k$ and $\mathbf{S}_k$ are the innovation mean and covariance evaluated by the filter update step. The partial derivatives of the negative log marginal likelihood function requires derivatives of the entire filtering recursion to be calculated. These rather lengthy equations are included in the supplementary material. The computational cost scales as $\mathcal{O}(nm^3)$, which makes this very beneficial if $m \ll n$.



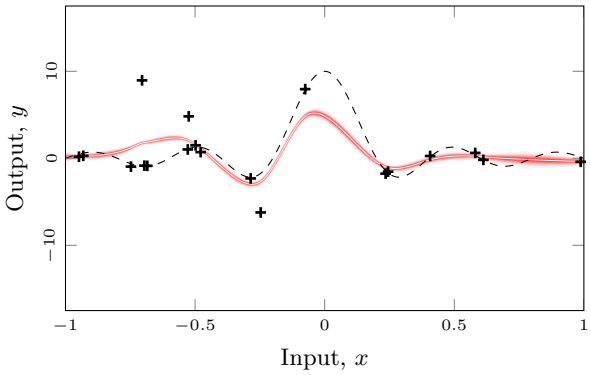
(a) Student-t filtering solution



(b) Student-t smoothing solution



(c) Gaussian filtering solution



(d) Gaussian smoothing solution

Figure 2: Comparison between the Student-t filtering and smoothing and the corresponding (Gaussian) Kalman filtering and Rauch–Tung–Striebel smoothing results for a set of measurements (+) of the scaled sinc function (- - -) featuring outliers. The shaded regions represent the 95% credible intervals. The hyperparameters were fixed to the same values for both the models.

Remark 3.5. *It is apparent that the filtering and smoothing scheme reverts to the Kalman filter and Rauch–Tung–Striebel smoother [6], as $\nu \to \infty$ (meaning $\gamma_k \to 1$, for $k = 1, 2, \ldots, n$).*

Figure 2 demonstrates the difference between the Gaussian (Kalman) and Student-t filter. Given fixed hyperparameters, the estimated means are equal, but the state covariances are not. This becomes apparent at the fourth observation (a clear outlier) at which the Student-t filter uncertainty grows to match that of the observation. The smoothing outcomes correspond exactly to the naive GP and TP predictions (see Sec. 2.1).

4 EXPERIMENTS

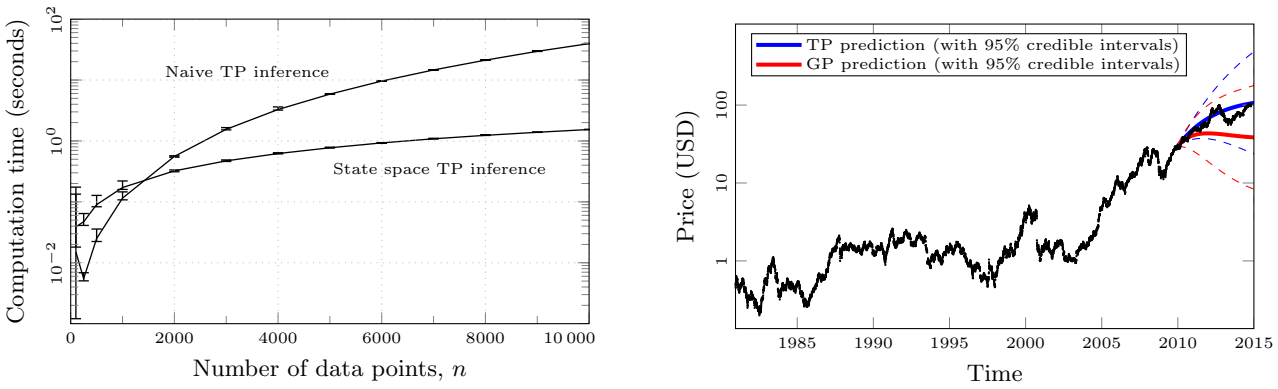
We demonstrate that the added flexibility and computational benefits make the state space inference scheme an appealing method for doing inference in data-intensive applications with long (or unbounded) time horizons.

4.1 Computational Efficiency

We illustrate the efficiency of the proposed inference scheme. We simulate data from a Gaussian process with a Matérn covariance function (smoothness 3/2; for the state space representation, see [3]) and corrupt the observations by independent Student-t noise. The state space solution is benchmarked against a naive TP implementation in Matlab (implemented as given in Sec. 2.1 using the techniques from [1]). Figure 3a shows the results for simulated TP regression problems with the number of observations ranging up to $n = 10\,000$. The empirical results agree with the theoretical derivations, and the computation time grows as $\mathcal{O}(n)$ for the state space and as $\mathcal{O}(n^3)$ for a naive implementation. Within numerical precision, both schemes returned identical results.

4.2 Comparisons Between GP and TP Regression via Synthetic Data

The following synthetic sets of data were considered. We sample 100 functions from a GP prior with a



(a) Demonstration of computational complexity

(b) Prediction in the stock price data

Figure 3: Demonstration of the computational benefits of the state space model in solving a TP regression problem. The error bars show the absolute minimum/maximum over the repetitions. The right-side figure shows predictions for a TP and GP after respective hyperparameter optimizations in the stock price data.

Matérn (smoothness 3/2) covariance function, and corrupt the observations by independent Gaussian noise (Synth A), independent Student-t noise (Synth B), and independent Gaussian noise with 25% of the observations being outliers with 100× larger noise variance (Synth C). Following [10], for each function we train on 80 observations and test on 20.

Table 1 shows the mean squared error and log likelihood for GP and TP regression results where the hyperparameters (along with σ_{n}^2, and ν for the TP) were optimized with respect to marginal likelihood using a conjugate-gradient optimizer. For added Gaussian noise (Synth A), all methods returned equal results, whereas for the long-tailed noises in (Synth B and Synth C) the TP model outperforms the GP (the GP failing to capture the latent function gives large standard deviations). The negligible differences between the state space and naive results are due to numerical issues along the hyperparameter optimization.

4.3 Household Electricity Consumption

The Student-t process can provide robustness to real-world inference problems, where the data is inherently noisy and corrupted by outlying observations. Observations of electricity consumption for one household (in kilowatt, ranging between 0.12 and 6.6) were made hourly over a time-period of 1 442 days ($n = 34\,087$, with 545 missing observations). We use hourly averages calculated from the even larger original data set[1]. We consider a GP and TP model solved by the fast state space inference methods. Prior knowledge of the daily and weekly rhythms are encoded into the GP/TP prior with the following quasi-periodic covariance structure:

$$k_{\boldsymbol{\theta}}(t,t') = k_{\mathrm{weekly}}(t,t')\, k_{\mathrm{Matérn}}(t,t') + k_{\mathrm{daily}}(t,t')\, k_{\mathrm{Matérn}}(t,t'), \quad (11)$$

where the periodic covariance functions (see, *e.g.*, [1]) have magnitude and length-scale hyperparameters, and the Matérn covariance (smoothness 5/2, unknown length-scales) allow the model to decay away from exact periodicity. This particular model is not suited for inducing point or basis function approximations (employed, *e.g.*, in [23]), but it has an approximate state space representation [22] (Taylor series truncated at 7 terms). We use 10-fold cross validation, with entire days left out for validation, and optimize the marginal likelihood with respect to all the 8 hyperparameters (including σ_{n}^2, and ν for the TP) in each fold.

As seen in Table 1, the TP gives a smaller error than the GP, which is primarily driven by the non-sensitivity to outliers during the hyperparameter optimization. Additionally, in this application the state space TP could be employed in real-time prediction.

4.4 Volatile Changes in Stock Price Data

As an example of a noisy regression problem we consider the stock market share price of Apple Inc. from December 2, 1980 onwards ($n = 8\,537$ trading days, see Fig. 3b). We model the log-price with a covariance function sum of a constant, linear, Matérn (smoothness 3/2), and exponential covariance function (see [1]). This is a non-stationary model, but it has an exact state space representation. We do 10-fold cross-validation on entire years left out for testing, and train all eight hyperparameters by maximizing the marginal likelihood. The results are included in Table 1.

[1]Data available from the UCI Machine Learning Repository: `http://mlr.cs.umass.edu/ml/datasets/Individual+household+electric+power+consumption`.

Table 1: GP and TP regression results computed by the direct naive solution and the state space methods. The table reports the log likelihood (LL) and mean squared error (MSE) for the test sets.

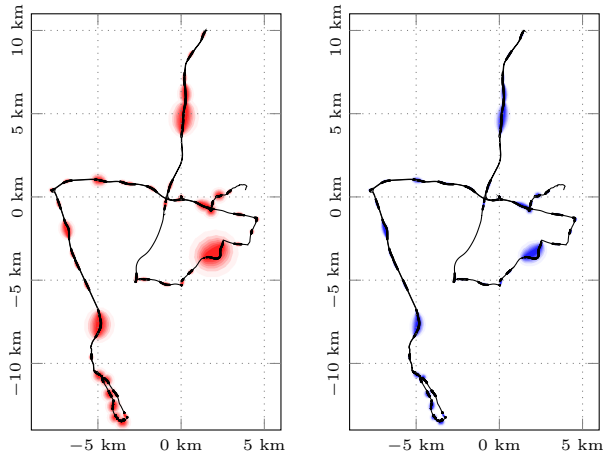
	Naive GP		State space GP		Naive TP		State space TP	
Data set	MSE	LL	MSE	LL	MSE	LL	MSE	LL
Synth A	$\mathbf{0.04 \pm 0.02}$	-25 ± 4	$\mathbf{0.04 \pm 0.02}$	-25 ± 4	$\mathbf{0.04 \pm 0.02}$	-25 ± 4	$\mathbf{0.04 \pm 0.02}$	-25 ± 4
Synth B	0.15 ± 0.28	-36 ± 17	0.13 ± 0.17	-36 ± 17	$\mathbf{0.12 \pm 0.11}$	-36 ± 17	$\mathbf{0.12 \pm 0.11}$	-36 ± 17
Synth C	0.56 ± 0.85	-46 ± 7	0.56 ± 0.85	-46 ± 7	$\mathbf{0.39 \pm 0.23}$	-46 ± 7	$\mathbf{0.38 \pm 0.21}$	-46 ± 7
Electricity	—	—	0.64 ± 0.20	-3261 ± 241	—	—	$\mathbf{0.56 \pm 0.04}$	-3287 ± 101
Stock	—	—	5.16 ± 15	-1700 ± 339	—	—	$\mathbf{0.41 \pm 0.84}$	-1701 ± 339

4.5 Interpolating GPS Location Data

Gaussian filtering is often employed in tracking applications, and the Student-t equivalent can help in quantifying the uncertainty caused by unreliable or missing observations. We consider a set of GPS data[2] marking the coordinates of a moving vehicle. The total size of the data set is $n = 6\,373$ observations collected over a time period of 106 minutes (sampling rate ~ 1 Hz).

We use a Wiener velocity model (see, *e.g.*, [6]) for the position $\mathbf{f}(t) = (f_{\mathrm{x}}(t), f_{\mathrm{y}}(t))$, where the acceleration (second derivative) of the vehicle is modeled as white noise: $\ddot{f}_{\mathrm{x}}(t) = w_{\mathrm{x}}(t)$ and $\ddot{f}_{\mathrm{y}}(t) = w_{\mathrm{y}}(t)$, where the two white noise processes share a common spectral density hyperparameter. We train a GP and TP model (optimize Q_{c}, σ_{n}^2, and ν w.r.t. marginal likelihood) off-line using the first five minutes of GPS data (302 observations).

For testing, we split the data randomly into batches of 30 seconds. Figure 4 shows the data with every third batch left out. The results from 10-fold cross-validation gives the Gaussian model an average MSE of 1589 versus 908 for the Student-t model. This difference stems primarily from the difference in the learned hyperparameter values, which is also apparent from the results in Figure 4.



(a) Gaussian (b) Student-t

Figure 4: 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%.

5 CONCLUSION

We have introduced a computationally efficient Bayesian filtering and smoothing based solution for inference in Student-t process (TP) regression models based on the entangled TP model formulation of [10], and an extension of the state space GP approach of [5] to TPs. The advantage of the approach is that the resulting Bayesian filtering and smoothing solution as well as the marginal likelihood evaluation can be implemented as a closed-form recursion which scales linearly (as opposed to cubicly) in the number of measurements. We have also demonstrated the practical computational benefits of the approach, and applied the method to synthetic and real-data examples.

The ideas presented in this paper can be extended in various ways. Spatio-temporal Student-t processes can be tackled with a similar approach as in [5], which results in stochastic partial differential equations with spatio-temporal Student-t process solutions. The resulting inference scheme is an infinite-dimensional generalization of the Bayesian filtering and smoothing scheme presented here. The state space representations of Student-t processes can also be combined with ordinary or partial differential equation based latent force models (LFMs, [24, 25]) in a computationally efficient state space form [26, 27]. Provided that the differential equations are linear, the inference can still be done by closed-form expressions.

An example Matlab implementation of the proposed method is available on the author web page: `http://arno.solin.fi`.

Acknowledgments

We thank Aki Vehtari for helpful discussions, and Tomi Peltola and Juho Kokkala for comments on the manuscript.

[2] Data available from the author web page.

References

[1] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[2] Mohinder S. Grewal and Angus P. Andrews. *Kalman Filtering: Theory and Practice Using MATLAB*. Wiley-Interscience, second edition, 2001.

[3] Jouni Hartikainen and Simo Särkkä. Kalman filtering and smoothing solutions to temporal Gaussian process regression models. In *IEEE International Workshop on Machine Learning for Signal Processing*, 2010.

[4] Simo Särkkä and Jouni Hartikainen. Infinite-dimensional Kalman filtering approach to spatio-temporal Gaussian process regression. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, volume 22 of *JMLR W&CP*, pages 993–1001, 2012.

[5] Simo Särkkä, Arno Solin, and Jouni Hartikainen. Spatiotemporal learning via infinite-dimensional Bayesian filtering and smoothing. *IEEE Signal Processing Magazine*, 30(4):51–61, 2013.

[6] Simo Särkkä. *Bayesian Filtering and Smoothing*, volume 3 of *Institute of Mathematical Statistics Textbooks*. Cambridge University Press, 2013.

[7] Kai-Tai Fang, Samuel Kotz, and Kai Wang Ng. *Symmetric Multivariate and Related Distributions*, volume 36 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1990.

[8] Bronius Grigelionis. *Student's t-Distribution and Related Stochastic Processes*. Springer, 2013.

[9] Anthony O'Hagan, Marc C. Kennedy, and Jeremy E. Oakley. Uncertainty analysis and other inference tools for complex computer codes. In *Bayesian Statistics*, volume 6, pages 503–524. Oxford University Press, 1999.

[10] Amar Shah, Andrew Gordon Wilson, and Zoubin Ghahramani. Student-t processes as alternatives to Gaussian processes. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, volume 33 of *JMLR W&CP*, pages 877–885, 2014.

[11] Michael Roth, Emre Özkan, and Fredrik Gustafsson. A Student's t filter for heavy tailed process and measurement noise. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5770–5774, 2013.

[12] Robert Piché, Simo Särkkä, and Jouni Hartikainen. Recursive outlier-robust filtering and smoothing for nonlinear systems using the multivariate Student-t distribution. In *IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6, 2012.

[13] G. Agamennoni, J. I. Nieto, and E. Nebot. Approximate inference in state-space models with heavy-tailed noise. *IEEE Transactions on Signal Processing*, 60 (10):5024–5037, 2012.

[14] Aleksandr Y. Aravkin, Bradley M. Bell, James V. Burke, and Gianluigi Pillonetto. An ℓ_1-Laplace robust Kalman smoother. *IEEE Transactions on Automatic Control*, 56(12):2898–2911, 2011.

[15] Dan Simon. *Optimal State Estimation: Kalman, H_∞, and Nonlinear Approaches*. John Wiley & Sons, 2006.

[16] Pasi Jylänki, Jarno Vanhatalo, and Aki Vehtari. Robust Gaussian process regression with a Student-t likelihood. *Journal of Machine Learning Research*, 12: 3227–3257, 2011.

[17] Samuel Kotz and Saralees Nadarajah. *Multivariate t Distributions and Their Applications*. Cambridge University Press, 2004.

[18] Harry L. Van Trees. *Detection, Estimation, and Modulation Theory, Part II: Nonlinear Modulation Theory*. John Wiley & Sons, New York, 1968–1971.

[19] Brian D.O. Anderson, John B. Moore, and Sonny G. Loo. Spectral factorization of time-varying covariance functions. *IEEE Transactions on Information Theory*, 15(5):550–557, 1969.

[20] Simo Särkkä and Robert Piché. On convergence and accuracy of state-space approximations of squared exponential covariance functions. In *Proceedings of IEEE International Workshop on Machine Learning for Signal Processing*, 2014.

[21] Arno Solin and Simo Särkkä. Gaussian quadratures for state space approximation of scale mixtures of squared exponential covariance functions. In *Proceedings of IEEE International Workshop on Machine Learning for Signal Processing*, 2014.

[22] Arno Solin and Simo Särkkä. Explicit link between periodic covariance functions and state space models. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, volume 33 of *JMLR W&CP*, pages 904–912, 2014.

[23] Thomas Nickson, Michael A. Osborne, Steven Reece, and Stephen J. Roberts. Automated machine learning on big data using stochastic algorithm tuning. *arXiv preprint arXiv:1407.7969*, 2014.

[24] Mauricio Álvarez and Neil D. Lawrence. Latent force models. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, volume 5 of *JMLR W&CP*, pages 9–16, 2009.

[25] Mauricio Álvarez, Jan R. Peters, Neil D. Lawrence, and Bernhard Schölkopf. Switched latent force models for movement segmentation. In *Advances in Neural Information Processing Systems*, volume 23, pages 55–63, 2010.

[26] Jouni Hartikainen and Simo Särkkä. Sequential inference for latent force models. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, 2011.

[27] Jouni Hartikainen, Mari Seppänen, and Simo Särkkä. State-space inference for non-linear latent force models with application to satellite orbit prediction. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

Supplementary Material

This is the supplementary material for '*State space methods for efficient inference in Student-t process regression*' by Solin and Särkkä published in Proceedings of the 18$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS). The references in this document point to the bibliography in the article.

1.1 Proof of Lemma 2.2

Proof. Let $\gamma \sim \mathrm{IG}(\alpha, \beta)$ be inverse gamma distributed with parameters α and β and $\mathbf{y} \mid \gamma \sim \mathrm{N}(\boldsymbol{\mu}, \gamma\mathbf{K})$. The scale mixture form of the probability density function can be written as

$$p(\mathbf{y}) = \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} \gamma^{-\alpha-1} \exp\left(-\frac{\beta}{\gamma}\right) \frac{1}{(2\pi)^{\frac{n}{2}}} \frac{1}{|\gamma\mathbf{K}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\frac{\Delta^2}{\gamma}\right) \mathrm{d}\gamma \tag{12}$$

$$= \frac{1}{\Gamma(\alpha)} \frac{1}{(2\beta\pi)^{\frac{n}{2}}} \frac{1}{|\mathbf{K}|^{\frac{1}{2}}} \int_0^\infty \xi^{\alpha+\frac{n}{2}-1} \exp\left(-\xi\left(1+\frac{\Delta^2}{2\beta}\right)\right) \mathrm{d}\xi \tag{13}$$

$$= \frac{\Gamma(\alpha+\frac{n}{2})}{\Gamma(\alpha)} \frac{1}{(2\beta\pi)^{\frac{n}{2}}} \frac{1}{|\mathbf{K}|^{\frac{1}{2}}} \left(1+\frac{\Delta^2}{2\beta}\right)^{-\left(\alpha+\frac{n}{2}\right)}, \tag{14}$$

where $\Delta^2 = (\mathbf{y}-\boldsymbol{\mu})^\mathsf{T}\mathbf{K}^{-1}(\mathbf{y}-\boldsymbol{\mu})$. We now recognize this as the Student-t density in Definition 2.1 by parametrizing $\alpha = \frac{\nu}{2}$ and $\beta = \frac{\nu-2}{2}$. Thus $\mathbf{y} \sim \mathrm{MVT}(\boldsymbol{\mu}, \mathbf{K}, \nu)$. Note the redundancy in $\gamma \sim \mathrm{IG}(\frac{\nu}{2}, \rho\frac{\nu-2}{2})$ and $\mathbf{y} \mid \gamma \sim \mathrm{N}(\mu, \frac{\gamma}{\rho}\mathbf{K})$ for $\rho > 0$. Without loss of generality, we choose $\rho = 1$. □

1.2 Marginal likelihood for the naive TP

We write down the negative log marginal likelihood (energy) function and its derivatives with respect to the degrees of freedom ν and the covariance hyperparameters $\boldsymbol{\theta} = (\sigma_\mathrm{n}^2, \theta_1, \theta_2, \ldots)$. The negative log marginal likelihood, $\mathcal{L} = -\log p(\mathbf{y} \mid \nu, \boldsymbol{\theta})$, is given by

$$\mathcal{L} = \frac{n}{2}\log((\nu-2)\pi) + \frac{1}{2}\log(|\mathbf{K}_{\boldsymbol{\theta}}|) - \log\left(\Gamma\left(\frac{\nu+n}{2}\right)\right) + \log\left(\Gamma\left(\frac{\nu}{2}\right)\right) + \frac{\nu+n}{2}\log\left(1+\frac{\beta}{\nu-2}\right), \tag{15}$$

where $\beta = \mathbf{y}^\mathsf{T}\mathbf{K}_{\boldsymbol{\theta}}^{-1}\mathbf{y}$. The derivatives can now be given as

$$\frac{\partial}{\partial\nu}\mathcal{L} = \frac{1}{2}\frac{n}{\nu-2} - \frac{1}{2}\psi\left(\frac{\nu+n}{2}\right) + \frac{1}{2}\psi\left(\frac{\nu}{2}\right) + \frac{1}{2}\log\left(1+\frac{\beta}{\nu-2}\right) - \frac{1}{2}\frac{(\nu+n)\beta}{(\nu-2)(\nu-2+\beta)}, \tag{16}$$

$$\frac{\partial}{\partial\theta_i}\mathcal{L} = \frac{1}{2}\mathrm{Tr}\left(\mathbf{K}_{\boldsymbol{\theta}}^{-1}\frac{\partial\mathbf{K}_{\boldsymbol{\theta}}}{\partial\theta_i}\right) + \frac{1}{2}\frac{\nu+n}{\nu-2+\beta}\mathbf{y}^\mathsf{T}\mathbf{K}_{\boldsymbol{\theta}}^{-1}\frac{\partial\mathbf{K}_{\boldsymbol{\theta}}}{\partial\theta_i}\mathbf{K}_{\boldsymbol{\theta}}^{-1}\mathbf{y}, \tag{17}$$

where $\psi(\cdot)$ is the digamma function.

1.3 Marginal likelihood for the state space TP

The negative log marginal likelihood can be evaluated recursively starting from $\mathcal{L}_0 = 0$:

$$\mathcal{L}_k = \mathcal{L}_{k-1} + \frac{1}{2}\log((\nu-2)\pi) + \frac{1}{2}\log(|\mathbf{S}_k|) + \log\Gamma\left(\frac{\nu_{k-1}}{2}\right) - \log\Gamma\left(\frac{\nu_k}{2}\right) + \frac{1}{2}\log\left(\frac{\nu_{k-1}-2}{\nu-2}\right) + \frac{\nu_k}{2}\log\left(1 + \frac{\mathbf{v}_k^\mathsf{T}\mathbf{S}_k^{-1}\mathbf{v}_k}{\nu_{k-1}-2}\right), \quad (18)$$

where $\mathbf{v}_k$ and $\mathbf{S}_k$ are the innovation mean and covariance evaluated by the filter update step, and $\nu_k = \nu_{k-1} + n_k$. Formally differentiating $\mathcal{L}_k$ gives a recursion algorithm for evaluating the gradient along with the filtering steps:

$$\frac{\partial\mathcal{L}_k(\boldsymbol{\theta})}{\partial\theta_i} = \frac{\partial\mathcal{L}_{k-1}(\boldsymbol{\theta})}{\partial\theta_i} + \frac{1}{2}\mathrm{Tr}\left(\mathbf{S}_k^{-1}(\boldsymbol{\theta})\,\frac{\partial\mathbf{S}_k(\boldsymbol{\theta})}{\partial\theta_i}\right) + \frac{\nu_k}{\nu_{k-1}-2+\mathbf{v}_k^\mathsf{T}\mathbf{S}_k^{-1}\mathbf{v}_k}\left(\mathbf{v}_k^\mathsf{T}(\boldsymbol{\theta})\,\mathbf{S}_k^{-1}(\boldsymbol{\theta})\,\frac{\partial\mathbf{v}_k(\boldsymbol{\theta})}{\partial\theta_i} - \frac{1}{2}\mathbf{v}_k^\mathsf{T}(\boldsymbol{\theta})\,\mathbf{S}_k^{-1}(\boldsymbol{\theta})\,\frac{\partial\mathbf{S}_k(\boldsymbol{\theta})}{\partial\theta_i}\,\mathbf{S}_k^{-1}(\boldsymbol{\theta})\,\mathbf{v}_k(\boldsymbol{\theta})\right). \quad (19)$$

The formal differentiation of the function also includes differentiating the filter prediction and update steps. This leads to the following rather lengthy recursion formulas, which include a lot of small matrix operations. On the filter prediction step we compute:

$$\frac{\partial\mathbf{m}_{k|k-1}(\boldsymbol{\theta})}{\partial\theta_i} = \frac{\partial\mathbf{A}_{k-1}(\boldsymbol{\theta})}{\partial\theta_i}\,\mathbf{m}_{k-1|k-1}(\boldsymbol{\theta}) + \mathbf{A}_{k-1}(\boldsymbol{\theta})\,\frac{\partial\mathbf{m}_{k-1|k-1}(\boldsymbol{\theta})}{\partial\theta_i}, \quad (20)$$

$$\frac{\partial\mathbf{P}_{k|k-1}(\boldsymbol{\theta})}{\partial\theta_i} = \frac{\partial\mathbf{A}_{k-1}(\boldsymbol{\theta})}{\partial\theta_i}\,\mathbf{P}_{k-1|k-1}(\boldsymbol{\theta})\,\mathbf{A}_{k-1}^\mathsf{T}(\boldsymbol{\theta}) + \mathbf{A}_{k-1}(\boldsymbol{\theta})\,\frac{\partial\mathbf{P}_{k-1|k-1}(\boldsymbol{\theta})}{\partial\theta_i}\,\mathbf{A}_{k-1}^\mathsf{T}(\boldsymbol{\theta}) + \mathbf{A}_{k-1}(\boldsymbol{\theta})\,\mathbf{P}_{k-1|k-1}(\boldsymbol{\theta})\,\frac{\partial\mathbf{A}_{k-1}^\mathsf{T}(\boldsymbol{\theta})}{\partial\theta_i} + \gamma_{k-1}(\boldsymbol{\theta})\frac{\partial\mathbf{Q}_{k-1}(\boldsymbol{\theta})}{\partial\theta_i} + \frac{\partial\gamma_{k-1}(\boldsymbol{\theta})}{\partial\theta_i}\mathbf{Q}_{k-1}(\boldsymbol{\theta}), \quad (21)$$

and on the filter update step we compute:

$$\frac{\partial\mathbf{v}_k(\boldsymbol{\theta})}{\partial\theta_i} = -\mathbf{H}\,\frac{\partial\mathbf{m}_{k|k-1}(\boldsymbol{\theta})}{\partial\theta_i}, \quad (22)$$

$$\frac{\partial\mathbf{S}_k(\boldsymbol{\theta})}{\partial\theta_i} = \mathbf{H}\frac{\partial\mathbf{P}_{k|k-1}(\boldsymbol{\theta})}{\partial\theta_i}\,\mathbf{H}^\mathsf{T}, \quad (23)$$

$$\frac{\partial\mathbf{K}_k(\boldsymbol{\theta})}{\partial\theta_i} = \frac{\partial\mathbf{P}_{k|k-1}(\boldsymbol{\theta})}{\partial\theta_i}\,\mathbf{H}^\mathsf{T}\,\mathbf{S}_k^{-1}(\boldsymbol{\theta}) - \mathbf{P}_{k|k-1}(\boldsymbol{\theta})\,\mathbf{H}^\mathsf{T}\,\mathbf{S}_k^{-1}(\boldsymbol{\theta})\,\frac{\partial\mathbf{S}_k(\boldsymbol{\theta})}{\partial\theta_i}\,\mathbf{S}_k^{-1}(\boldsymbol{\theta}), \quad (24)$$

$$\frac{\partial\mathbf{m}_{k|k}(\boldsymbol{\theta})}{\partial\theta_i} = \frac{\partial\mathbf{m}_{k|k-1}(\boldsymbol{\theta})}{\partial\theta_i} + \frac{\partial\mathbf{K}_k(\boldsymbol{\theta})}{\partial\theta_i}\,\mathbf{v}_k(\boldsymbol{\theta}) + \mathbf{K}_k(\boldsymbol{\theta})\,\frac{\partial\mathbf{v}_k(\boldsymbol{\theta})}{\partial\theta_i}, \quad (25)$$

$$\frac{\partial\mathbf{P}_{k|k}(\boldsymbol{\theta})}{\partial\theta_i} = \frac{\gamma_k(\boldsymbol{\theta})}{\gamma_{k-1}(\boldsymbol{\theta})}\left(\frac{\partial\mathbf{P}_{k|k-1}(\boldsymbol{\theta})}{\partial\theta_i} - \frac{\partial\mathbf{K}_k(\boldsymbol{\theta})}{\partial\theta_i}\,\mathbf{S}_k(\boldsymbol{\theta})\,\mathbf{K}_k^\mathsf{T}(\boldsymbol{\theta}) - \mathbf{K}_k(\boldsymbol{\theta})\,\frac{\partial\mathbf{S}_k(\boldsymbol{\theta})}{\partial\theta_i}\,\mathbf{K}_k^\mathsf{T}(\boldsymbol{\theta}) - \mathbf{K}_k(\boldsymbol{\theta})\,\mathbf{S}_k(\boldsymbol{\theta})\,\frac{\partial\mathbf{K}_k^\mathsf{T}(\boldsymbol{\theta})}{\partial\theta_i}\right) + \frac{1}{\gamma_{k-1}(\boldsymbol{\theta})}\left(\frac{\partial\gamma_k(\boldsymbol{\theta})}{\partial\theta_i} - \frac{\gamma_k(\boldsymbol{\theta})}{\gamma_{k-1}(\boldsymbol{\theta})}\frac{\partial\gamma_{k-1}(\boldsymbol{\theta})}{\partial\theta_i}\right)\left(\mathbf{P}_{k|k-1} - \mathbf{K}_k\mathbf{S}_k\mathbf{K}_k^\mathsf{T}\right), \quad (26)$$

$$\frac{\partial\gamma_k(\boldsymbol{\theta})}{\partial\theta_i} = \frac{\partial\gamma_{k-1}(\boldsymbol{\theta})}{\partial\theta_i}\,\frac{\nu_{k-1}-2+\mathbf{v}_k^\mathsf{T}\mathbf{S}_k^{-1}\mathbf{v}_k}{\nu_k-2} + \frac{\gamma_{k-1}(\boldsymbol{\theta})}{\nu_k-2}\left(2\,\mathbf{v}_k^\mathsf{T}(\boldsymbol{\theta})\,\mathbf{S}_k^{-1}(\boldsymbol{\theta})\,\frac{\partial\mathbf{v}_k(\boldsymbol{\theta})}{\partial\theta_i} - \mathbf{v}_k^\mathsf{T}(\boldsymbol{\theta})\,\mathbf{S}_k^{-1}(\boldsymbol{\theta})\,\frac{\partial\mathbf{S}_k(\boldsymbol{\theta})}{\partial\theta_i}\,\mathbf{S}_k^{-1}(\boldsymbol{\theta})\,\mathbf{v}_k(\boldsymbol{\theta})\right). \quad (27)$$

Note that, the derivative $\frac{\partial\mathcal{L}}{\partial\nu}$ can be evaluated as given in Equation (16), if the $\beta = \beta_n$ is evaluated along the filtering recursion such that $\beta_k = \beta_{k-1} + \gamma_{k-1}\mathbf{v}_k^\mathsf{T}\mathbf{S}_k^{-1}\mathbf{v}_k$ and starting from $\beta_0 = 0$. For maximum *a posteriori* estimation, the recursion should be started from the initial condition $\frac{\partial\mathcal{L}_0(\boldsymbol{\theta})}{\partial\theta_i} = -\frac{\partial\log p(\boldsymbol{\theta})}{\partial\theta_i}$. For a similar formulation for the Gaussian filter, see [6] and the references therein.