KALMAN FILTERING AND SMOOTHING SOLUTIONS TO TEMPORAL GAUSSIAN PROCESS REGRESSION MODELS

Jouni Hartikainen and Simo Särkkä

Aalto University
Department of Biomedical Engineering and Computational Science
Rakentajanaukio 2, 02150 Espoo, Finland

ABSTRACT

In this paper, we show how temporal (i.e., time-series) Gaussian process regression models in machine learning can be reformulated as linear-Gaussian state space models, which can be solved exactly with classical Kalman filtering theory. The result is an efficient non-parametric learning algorithm, whose computational complexity grows linearly with respect to number of observations. We show how the reformulation can be done for Matérn family of covariance functions analytically and for squared exponential covariance function by applying spectral Taylor series approximation. Advantages of the proposed approach are illustrated with two numerical experiments.

1. INTRODUCTION

In this paper, we shall show how a wide class of temporal (i.e., time-series) Gaussian process regression problems can be reformulated as Kalman filtering and Rauch-Tung-Striebel (RTS) smoothing of linear state space models. The advantage of the Kalman filter and RTS smoother based approach is that the computational complexity grows as $\mathcal{O}(n)$, where n is the number of measurements. With direct Gaussian process regression methods the complexity is $\mathcal{O}(n^3)$, although in temporal regression context much work has been done (e.g., [1, 2]) to alleviate the computational burden of the direct GP solution by utilizing the structural form of the joint covariance matrix of the process (e.g., by representing it in Toeplitz form).

In order to apply the Kalman filters and smoothers, the model has to be reformulated as estimation of the state of a multi-dimensional continuous-time Gauss-Markov process. We shall show how for Matérn class of covariance functions this reformulation can be done without an approximation and how a simple spectral Taylor series approximation can be used for squared exponential covariance functions. This approach is directly applicable to cases with missing or non-uniformly sampled data, which is a typical problem for methods utilizing the temporal structure of data in covariance representations.

The underlying idea of efficient inference of Gaussian processes using a state space formulation is not new, because the contribution of Kalman's original 1960's article [3] was exactly this kind of re-formulation of the filtering problem of Wiener [4]. The idea of Bayesian modeling of unknown processes as Gaussian processes has also been widely utilized in communications theory [5, 6], but the philosophy differs slightly from that of Gaussian process regression in machine learning [7]. The idea of approximating the squared exponential covariance function with spectral Taylor series expansion is also not new, and can be found, for example, in [8].

Kalman filtering and Rauch-Tung-Striebel smoothing are also mature subjects, and the solutions to discrete time and continuous time linear Gaussian state space models (Gaussian processes) were first presented in articles [3, 9, 10]. Although the articles talk in language of least squares and maximum likelihood estimates, the results are completely equivalent to the full Bayesian treatment of the problems (see, e.g. [5, 6, 11]). This connection of Gaussian process regression to Kalman filtering and other fields was also discussed in the discussion part of O'Hagan's 1978 article [12].

In this paper we shall apply some these useful classical results in machine learning context to derive efficient learning algorithms for temporal Gaussian process regression problems.

2. GAUSSIAN PROCESS REGRESSION

Gaussian process (GP) regression [12, 7] concerns the problem of estimating the value of a function $f : \Re^D \mapsto \Re$ on arbitrary input point $\mathbf{t}_* \in \Re^D$, given a set of data $\mathcal{D} = \{\mathbf{t}_i, y_i\}_{i=1}^n$, where the training targets $y_i \in \Re$ are usually assumed to be the function values corrupted with Gaussian

JH thanks Finnish Graduate School in Computational Sciences (FICS) for funding. SS thanks Academy of Finland's Centre of Excellence in Computational Complex Systems Research for funding. The authors would like to thank Aki Vehtari for helpful discussion and support during the work.

noise:

$$y_i = f(\mathbf{t}_i) + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2_{\text{noise}}).$$

The non-parametric GP approach is based on assuming that the joint prior for values of f on any collection of points $\mathbf{t}$ is Gaussian, commonly denoted as

$$f(\mathbf{t}) \sim \mathcal{GP}(0, k(\mathbf{t}, \mathbf{t}', \theta)),$$

where k is a covariance function with hyperparameters θ. This means that the prior for the function values on a finite set of input points $\{\mathbf{t}_1 \cdots \mathbf{t}_n\}$ is $p(f(\mathbf{t}_1), \ldots, f(\mathbf{t}_n)) \sim N(\mathbf{0}, \mathbf{K})$, where the entries of the covariance matrix $\mathbf{K}$ are formed by evaluating the covariance function as $[\mathbf{K}]_{ij} = k(\mathbf{t}_i, \mathbf{t}_j, \theta)$. Due to the well known properties of the Gaussian distribution it is straightforward to show that given the data set $\mathcal{D}$ and hyperparameters θ, the posterior of f on input $\mathbf{t}_*$ is also Gaussian

$$p(f(\mathbf{t}_*)|\mathbf{y}, \theta) = N(\mu_{\mathcal{GP}}(\mathbf{t}_*), \sigma^2_{\mathcal{GP}}(\mathbf{t}_*)),$$

with mean and variance

$$\mu_{\mathcal{GP}}(\mathbf{t}_*) = \mathbf{k}_{*,f}(\mathbf{K} + \sigma^2_{\text{noise}}\mathbf{I})^{-1}\mathbf{y}$$
$$\sigma^2_{\mathcal{GP}}(\mathbf{t}_*) = k(\mathbf{t}_*, \mathbf{t}_*) - \mathbf{k}_{*,f}(\mathbf{K} + \sigma^2_{\text{noise}}\mathbf{I})^{-1}\mathbf{k}^T_{*,f},$$

where the element i of row vector $\mathbf{k}_{*,f}$ is the prior covariance between $f(\mathbf{t}_*)$ and $f(\mathbf{t}_i)$ as $[\mathbf{k}_{*,f}]_i = k(\mathbf{t}_*, \mathbf{t}_i, \theta)$.

The central problem with the brute-force implementation of GP models are the $\mathcal{O}(n^3)$ and $\mathcal{O}(n^2)$ scalings of computational complexity and memory requirements, with respect to number of data points. In this paper we concentrate on time series data ($D = 1$) and stationary covariance functions (i.e., $k(t, t', \theta) = k(\tau, \theta)$, where $\tau = t - t'$), and show how the inference can be done in $\mathcal{O}(n)$ in such cases with the most commonly used classes of covariance functions. For notational compactness we omit the hyperparameters θ from the results below by treating them as fixed, and discuss their selection briefly in Section 5.

3. SPECTRA AND COVARIANCE FUNCTIONS OF STOCHASTIC DIFFERENTIAL EQUATIONS

In this article we aim to represent the random process $f(t)$ having the covariance function $k(\tau)$ as output of a linear time invariant (LTI) stochastic differential equation (SDE), which are efficient to analyse with Kalman filtering theory discussed in Section 5. In particular, we consider mth order scalar LTI SDEs of form

$$\frac{d^m f(t)}{dt^m} + a_{m-1}\frac{d^{m-1} f(t)}{dt^{m-1}} + \cdots + a_1 \frac{df(t)}{dt} + a_0 f(t) = w(t), \tag{1}$$

where $a_0, \ldots, a_{m-1}$ are known constants and $w(t)$ is a white noise process with spectral density $S_w(\omega) = q$. This can be written as first order (vector) Markov process[1]

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{F}\,\mathbf{x}(t) + \mathbf{L}\,w(t), \tag{2}$$

where the state $\mathbf{x}(t)$ contains the derivatives of $f(t)$ up to order $m-1$ as $\mathbf{x}(t) = (f(t)\ \frac{df(t)}{dt}\ \cdots\ \frac{d^{m-1}f(t)}{dt^{m-1}})^T$. The matrices $\mathbf{F} \in \Re^{m\times m}$ and $\mathbf{L} \in \Re^{m\times 1}$ can be written as

$$\mathbf{F} = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ -a_0 & \cdots & -a_{m-2} & -a_{m-1} \end{pmatrix},\ \mathbf{L} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}.$$

This representation is known as the companion form [13] of (1), and there exists other forms as well (such as the canonical controllable and observable forms [14]), which could be more stable numerically, should some problems arise during the inference. With LTI SDEs of orders considered in this paper, however, we did not experience any problems.

By defining $\mathbf{H} = (1\ 0\ \cdots\ 0)^T$ we can extract $f(t)$ from $\mathbf{x}(t)$ as $f(t) = \mathbf{H}\,\mathbf{x}(t)$. This can be used to compute the power spectral density of $f(t)$ by replacing $\mathbf{x}(t)$ with $\mathbf{H}\,\mathbf{x}(t)$ in Equation (2) and formally taking the Fourier transform on both sides of it, which after some algebra yields

$$S(\omega) = \mathbf{H}(\mathbf{F} + i\omega\mathbf{I})^{-1}\mathbf{L}\,q\,\mathbf{L}^T[(\mathbf{F} - i\omega\mathbf{I})^{-1}]^T\mathbf{H}^T. \tag{3}$$

In stationary state (i.e. when the process has run an infinite amount of time) the covariance function of $f(t)$ is the inverse Fourier transform of its spectral density:

$$k(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S(\omega)e^{i\omega\tau}d\omega.$$

This can be calculated more easily as [6]

$$k(\tau) = \begin{cases} \mathbf{H}\mathbf{P}_\infty\Phi(\tau)^T\mathbf{H}^T, & \text{if } \tau \geq 0 \\ \mathbf{H}\Phi(-\tau)\mathbf{P}_\infty\mathbf{H}^T, & \text{if } \tau < 0, \end{cases} \tag{4}$$

where $\Phi(\tau) = \exp(\mathbf{F}\,\tau)$ and $\mathbf{P}_\infty$ is the stationary covariance of $\mathbf{x}(t)$. The latter can be obtained as the solution of the matrix Riccati equation

$$\frac{d\mathbf{P}}{dt} = \mathbf{F}\,\mathbf{P} + \mathbf{P}\,\mathbf{F}^T + \mathbf{L}\,q\,\mathbf{L}^T = 0. \tag{5}$$

4. CONVERTING COVARIANCE FUNCTIONS TO STOCHASTIC DIFFERENTIAL EQUATIONS

Assume now that we have been given a stationary covariance function $k(\tau)$ for $f(t)$, and we wish to represent $f(t)$

[1] Note that here we use the white-noise notation for SDEs, whereas the corresponding formal Ito SDE notation would be $d\mathbf{x} = \mathbf{F}\,\mathbf{x}(t)\,dt + \mathbf{L}\,d\beta(t)$, where $\beta(t)$ is a Brownian motion.

in form (2) to achieve the linear time inference. The real question now is how to form $\mathbf{F}$, $\mathbf{L}$ and q such that the first component of $\mathbf{x}(t)$ has the desired covariance function $k(\tau)$. This is possible to do for covariance functions, whose spectral density $S(\omega)$ can be represented as a rational function of the form

$$S(\omega) = \frac{(\text{constant})}{(\text{polynomial in } \omega^2)}, \tag{6}$$

which is in fact the functional form of (3). By applying spectral factorization [4, 5, 15] we can write the spectral density as

$$S(\omega) = H(\mathrm{i}\omega)\, q\, H(-\mathrm{i}\omega), \tag{7}$$

where the transfer functions $H(\mathrm{i}\omega)$ and $H(-\mathrm{i}\omega)$ have all of their poles in upper and lower planes, respectively. We can construct a stable (causal) Markov process with the former, which means that when a white noise process with spectral density q is fed as input to the system with transfer function $H(i\omega)$, the output has the desired spectral density. This leads to following frequency domain representation of the process:

$$(i\omega)^m F(\omega) + h_{m-1}(i\omega)^{m-1} F(\omega) \ \cdots + h_0 F(\omega) = W(\omega),$$

where $W(\omega)$ and $F(\omega)$ are the formal Fourier transforms of $w(t)$ and $f(t)$, and $h_0, \ldots, h_{m-1}$ the coefficients of polynomial in the denominator of $H(\mathrm{i}\omega)$. In time domain:

$$\frac{d^m f(t)}{dt^m} + h_{m-1}\frac{d^{m-1} f(t)}{dt^{m-1}} + \cdots + h_1 \frac{df(t)}{dt} + h_0 f(t) = w(t),$$

which is of the desired Markov form of Equation (1). Next we show how some of the most widely used classes of covariance functions can be transformed to models of this form.

4.1. Whittle-Matérn Family

A common class of covariance functions is the Whittle-Matérn family [16, 17, 7]

$$k_\nu(\tau) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}}{l}\tau \right)^\nu K_\nu \left(\frac{\sqrt{2\nu}}{l}\tau \right), \tag{8}$$

where l and σ^2 are the length scale and magnitude hyperparameters controlling the overall correlation scale and variability of the process, K_ν is a modified Bessel-function of the second kind and ν a parameter controlling the smoothness of the process. With $\nu = 1/2$ this is equivalent with the exponential covariance $k(\tau) = \sigma^2 \exp(-|\tau|/l)$, and when $\nu \to \infty$ we obtain the squared exponential (see Section 4.2). This parametrization is the same as in the book of Rasmussen and Williams [7]. With one-dimensional processes the spectral density of the Matérn covariance function (8) is

$$S(\omega) = \sigma^2 \frac{2\pi^{1/2}\Gamma(\nu + 1/2)}{\Gamma(\nu)} \lambda^{2\nu} (\lambda^2 + \omega^2)^{-(\nu+1/2)},$$

where we have denoted $\lambda = \sqrt{2\nu}/l$. In this paper we limit our view to cases in which $\nu = p + 1/2$, where p is a non-negative integer. Thus,

$$S(\omega) \propto (\lambda^2 + \omega^2)^{-(p+1)}.$$

Clearly this is of the desired rational function form (6), and no approximation is needed. We can rewrite this as

$$S(\omega) \propto (\lambda + \mathrm{i}\omega)^{-(p+1)} (\lambda - \mathrm{i}\omega)^{-(p+1)},$$

from which we can extract the transfer function of a stable Markov process as

$$H(\mathrm{i}\omega) = (\lambda + \mathrm{i}\omega)^{-(p+1)}.$$

The corresponding spectral density of the white noise process $w(t)$ is

$$q = \frac{2\sigma^2 \pi^{1/2} \lambda^{(2p+1)} \Gamma(p+1)}{\Gamma(p+1/2)}. \tag{9}$$

For example, with $p = 1$ the corresponding LTI SDE model reads

$$\frac{d\mathbf{x}(t)}{dt} = \begin{pmatrix} 0 & 1 \\ -\lambda^2 & -2\lambda \end{pmatrix} \mathbf{x}(t) + \begin{pmatrix} 0 \\ 1 \end{pmatrix} w(t),$$

and with $p = 2$

$$\frac{d\mathbf{x}(t)}{dt} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -\lambda^3 & -3\lambda^2 & -3\lambda \end{pmatrix} \mathbf{x}(t) + \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} w(t),$$

where in both cases the spectral density q of white noise process $w(t)$ is evaluated by Equation (9).

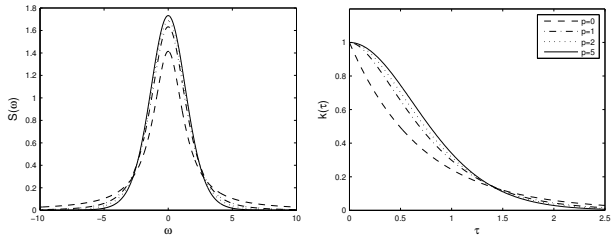
The spectral densities and covariance functions for constructed LTI SDE models with $p \in \{0, 1, 2, 5\}$ evaluated by equations (3) and (4) are shown in Figure 1. It can be seen that the Matérn family of covariance functions provides a flexible way of controlling the smoothness of the process.

4.2. Squared Exponential

A very commonly used covariance function in machine learning setting is the squared exponential covariance function

$$k(\tau) = \sigma^2 \exp\left(-\frac{\tau^2}{2l^2} \right), \tag{10}$$

where the length scale and magnitude hyperparameters l and σ^2 have the same interpretation as with the Matérn covariance function (8). A process $f(t)$ with covariance function (10) is infinitely differentiable, which means that there does not exists a finite-dimensional Markov process having exactly the same spectral density as $f(t)$, but in this article we aim to find a finite-dimensional Markov process, which has approximately the same spectral density.



(a) $S(\omega)$ (b) $k(\tau)$

Fig. 1. Spectral density $S(\omega)$ and covariance function $k(\tau)$ of the Matérn covariance function model with parameter p having the values $p \in \{0, 1, 2, 5\}$.

By denoting $\kappa = \frac{1}{2l^2}$ the exact spectral density $S(\omega)$ of the process $f(t)$ with the squared exponential covariance function is

$$S(\omega) = \sigma^2 \sqrt{\frac{\pi}{\kappa}} \exp\left(-\frac{\omega^2}{4\kappa}\right).$$

If we think $S(\omega)$ as function of ω^2, we may form Taylor series approximation to $1/S(\omega)$ as follows:

$$\begin{aligned}
\frac{1}{S(\omega)} &= \frac{1}{\sigma^2}\sqrt{\frac{\kappa}{\pi}} \exp\left(\frac{\omega^2}{4\kappa}\right) \\
&\approx \frac{1}{\sigma^2}\sqrt{\frac{\kappa}{\pi}} \left(1 + \frac{\omega^2}{4\kappa} + \frac{1}{2!}\frac{\omega^4}{(4\kappa)^2} + \cdots + \frac{1}{N!}\frac{\omega^{2N}}{(4\kappa)^N}\right) \\
&= \frac{1}{\sigma^2 N!(4\kappa)^N}\sqrt{\frac{\kappa}{\pi}} \Big(N!(4\kappa)^N + N!(4\kappa)^{N-1}\omega^2 \\
&\quad + \frac{N!(4\kappa)^{N-2}}{2!}\omega^4 + \cdots + \omega^{2N}\Big)
\end{aligned}$$

To simplify the results, we shall assume that N is even. That is, the spectral density of the original process can be approximated by the following spectral density:

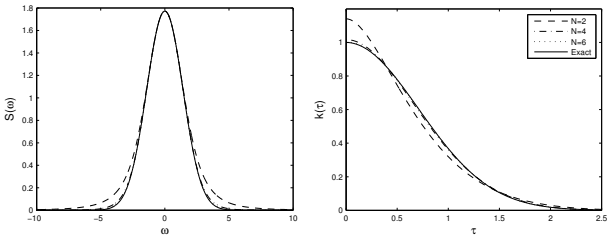
$$\hat{S}(\omega) = \sigma^2 N!(4\kappa)^N \sqrt{\frac{\pi}{\kappa}} \left(\frac{1}{\sum_{n=0}^{N} \frac{N!(4\kappa)^{N-n}}{n!}\omega^{2n}}\right),$$

which is spectral density of a finitely differentiable process, because it is a rational function of ω^2. With this we can now approximate the spectral density of the infinitely differentiable process $f(t)$.

In order to find the transfer function $H(\mathrm{i}\omega)$ in (7) we write denominator $P(\mathrm{i}\omega)$ of the spectral density $\hat{S}(\omega)$ as a polynomial of $\mathrm{i}\omega$:

$$P(\mathrm{i}\omega) = \sum_{n=0}^{N} \frac{N!(-1)^n(4\kappa)^{N-n}}{n!}(\mathrm{i}\omega)^{2n}, \qquad (11)$$

Because N is even, the coefficient of $(\mathrm{i}\omega)^{2N}$ is 1 and we can now form $H(\mathrm{i}\omega)$ as follows:



(a) $S(\omega)$ (b) $k(\tau)$

Fig. 2. Spectral density $S(\omega)$ and covariance function $k(\tau)$ of the approximate squared exponential covariance function model with orders $N \in \{2, 4, 6\}$. For comparison, values for the exact squared exponential are also shown, but it is hardly distinguishable from the approximation of order $N = 6$.

1. Compute the roots of the polynomial $P(x)$ defined in Equation (11). This can be done numerically with a computer program.

2. Construct two polynomials $P^-(x)$ and $P^+(x)$, where $P^-(x)$ has the roots of $P(x)$ with negative real parts and $P^+(x)$ has the roots with positive real parts and

$$P(x) = P^-(x)\, P^+(x).$$

3. The transfer function and the corresponding white noise specral density are now given as

$$H(\mathrm{i}\omega) = \frac{1}{P^-(\mathrm{i}\omega)}, \quad q = \sigma^2 N!(4\kappa)^N \sqrt{\frac{\pi}{\kappa}}.$$

This approximation was originally presented briefly in [8], but here we present its connection to GP regression models more explicitly and show how it can be a practical tool for inference.

The approximate spectral densities and covariance functions for $N \in \{2, 4, 6\}$ evaluated by equations (3) and (4) are shown in Figure 2. The spectral density around the origin is well approximated with all the presented orders, which is natural since the Taylor series approximation was formed around the origin. With $N = 2$ the tails of the density deviate from the true values, while with $N = 4$ there is only some difference and with $N = 6$ the approximate density cannot be easily distinguished from the exact value. With the covariance function the effect approximation is the opposite: with $N = 2$ the tail of the function is reasonably well approximated, but the near the origin there is a considerable offset from the true value. With $N = 4$ there is a noticeable deviation from the exact function, while with $N = 6$ the function is practically identical to the true one.

5. INFERENCE WITH STATE-SPACE MODELS

The continuous time LTI model (2) above can be transformed into discrete time model of the following form:

$$\mathbf{x}_k = \mathbf{A}_{k-1}\mathbf{x}_{k-1} + \mathbf{q}_{k-1}, \quad \mathbf{q}_{k-1} \sim N(\mathbf{0}, \mathbf{Q}_{k-1}),$$

where the state transition and process noise covariance matrices can be calculated analytically (see, e.g., [15]) as

$$\mathbf{A}_{k-1} = \Phi(\Delta t_k)$$
$$\mathbf{Q}_{k-1} = \int_0^{\Delta t_k} \Phi(\Delta t_k - \tau)\, \mathbf{L}\, q\, \mathbf{L}^T\, \Phi(\Delta t_k - \tau)^T \mathrm{d}\tau,$$

where $\Delta t_k = t_k - t_{k-1}$. The measurement model is of form

$$y_k = \mathbf{H}\mathbf{x}_k + \epsilon_k,\ \epsilon_k \sim N(0, \sigma^2_{\text{noise}}).$$

The posterior distribution of state trajectory $p(\mathbf{x}_{1:n}|y_{1:n})$ for this class of models can be exactly solved with the well-known Kalman filter [3] and Rauch-Tung-Striebel smoother [10] algorithms, which scale $\mathcal{O}(nm^3)$ and $\mathcal{O}(nm^2)$ in computational complexity and memory requirements. In this context, the state dimensionality m is typically very small (say, less than 10) and constant with respect to n, so the scalings are $\mathcal{O}(n)$ in practice. If one wishes to obtain the same result as with the brute-force naive GP implementation reviewed in Section 2, estimation with Kalman filter must be started from the prior $p(\mathbf{x}_0) = N(\mathbf{0}, \mathbf{P}_\infty)$, where $\mathbf{P}_\infty$ is the stationary covariance solving the Riccati equation (5). It is also useful to note that missing data can be treated by simply skipping the update step of Kalman filter, and non-uniformly sampled data by recomputing $\mathbf{A}_k$ and $\mathbf{Q}_k$ between the measurements.

In machine learning context it has become accustomary to choose the covariance function hyperparameters θ with GP models by optimizing the marginal data likelihood

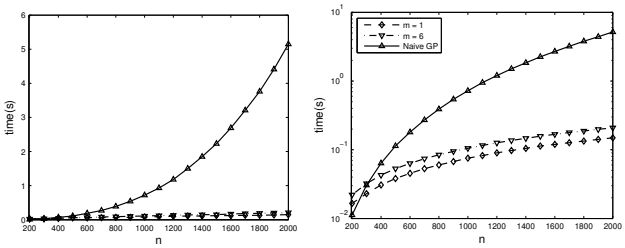
$$p(y_{1:n}|\theta) = \prod_{i=1}^{n} p(y_i|y_{1:i-1}, \theta).$$

These factors are computed as by-products of the Kalman filter algorithm, and thereby our approach naturally lends itself to marginal likelihood based learning of hyperparameters. The computations are especially efficient since smoothing pass is not even required when computing the marginal likelihood terms. It is also possible to recursively calculate analytic gradients of $p(y_i|y_{1:i-1}, \theta)$ with respect to θ, but due to space limitations we omit the details of the fairly lengthy equations here.

6. EXPERIMENTS

6.1. Comparison of Computational Complexity

To illustrate the efficiency of the proposed approach we calculated the CPU time needed for inference with both the proposed Kalman filtering method and the brute-force GP implementation of Section 2 with different data set sizes. The results are shown in Figure 3 (a) (normal scale) and (b) (log scale). It is obvious that the brute-force GP method is applicable only to moderately sized data sets (n is about few thousand) in a reasonable time frame, while the Kalman filter/smoother can be applied to practically endless streams of data. For example, with $m = 6$ and $n = 10^6$ the inference takes about 10 seconds and with $n = 10^7$ about 100 seconds [2].



(a) Normal scale (b) Log scale

Fig. 3. Time needed for inference for the Kalman filtering approach ($m \in \{0, 6\}$) and for the naive GP implementation as a function of data set size. Panel (a) shows the times in normal scale and panel (b) in log scale.

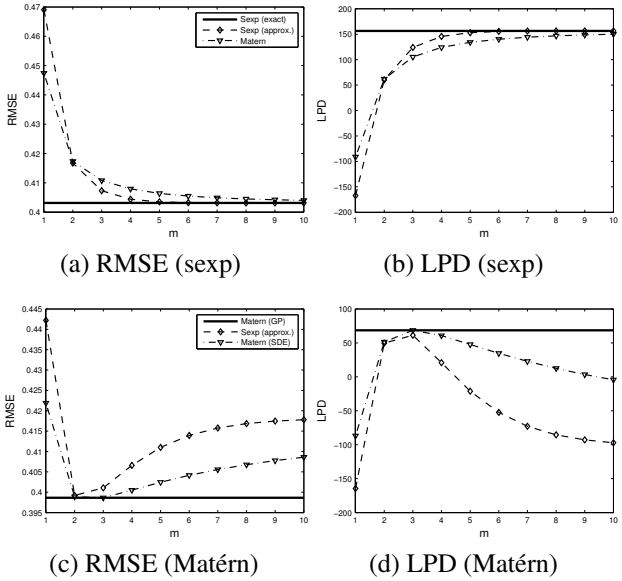
6.2. Effect of Approximation

In this experiment we investigate numerically how the approximation made for the squared exponential covariance function affects the regression results when compared to solution of a exact squared exponential model. We generated 100 time series with a GP having the squared exponential covariance function ($l = 10$ and $\sigma^2 = 1$) with each data set containing $n = 500$ data points. For 400 of these we created observations with the Gaussian noise model ($\sigma^2_{\text{noise}} = 0.1$), and treated all 500 as test points.

Figures 4 (a) and (b) show the root mean square error (RMSE) and log predictive density (LPD) values averaged over the 100 simulation runs for the squared exponential and Matérn LTI SDE models as well as for the exact GP. With all models the hyperparameters were fixed to true values. It can be seen that squared exponential models converge more quickly toward the exact squared exponential solution. This does not, however, mean that squared exponential is the better model in all situations as its hard to imagine a real physical process, which is infinitely differentiable.

As a second experiment we generated another set of 100 time series with the covariance function of the GP changed

[2] The software was written in Matlab and run on Intel Core 2 Quad 2.83 GHz, 8GB RAM desktop PC. The multi-threading abilities of Matlab were disabled in all of the experiments to make comparisons fair.



(a) RMSE (sexp) (b) LPD (sexp)

(c) RMSE (Matérn) (d) LPD (Matérn)

Fig. 4. Estimation of a processes with a exact squared exponential (panels (a) and (b)) and Matérn $\nu = 5/2$ (panels (c) and (d)) covariance functions. Panels show the RMSE and LPD values with squared exponential and Matérn LTI SDE models with orders $m = [1, \ldots, 10]$. In both cases values for the exact GP solutions are also shown.

to the Matérn covariance function with $\nu = 5/2$ (i.e. $p = 2$) and same hyperparameters as above. The corresponding results are shown Figures 4 (c) and (d). Clearly the SDE and GP versions of Matérn $\nu = 5/2$ covariance give identical results, which of course is to expected since there was no need to do any approximations for the spectral density of Matérn covariance function.

7. CONCLUSIONS

In this paper we have shown how GP regression for temporal data can be done in linear time with the most commonly used classes of covariance functions. The approach is based on reformulating the regression problem as a smoothing problem of a linear-Gaussian state space model. This is possible for covariance functions whose spectral density can be represented as a rational function of squared angular frequency. For Matérn family of covariance functions this can be done without any approximations, and for the frequently used squared exponential covariance by applying simple Taylor series approximation for the spectral density.

The resulting algorithm is very efficient, and can be applied to huge temporal data sets in a reasonable time frame. Furthermore, the approach lends itself directly to marginal likelihood based learning of hyperparameters, and handles missing and non-uniformly sampled data points without any problems. Although this view on Gaussian process models is not a new one, we expect it to gain more emphasis in future, since the advantages of the state space representation can be very substantial in many application areas.

8. REFERENCES

[1] Y. Zhang, W. E. Leithand, and D.J. Leith, "Time-series Gaussian process regression based on Toeplitz computation of $O(N^2)$ operations and $O(N)$-level storage," in *Proc. of the 44th IEEE Conf. on Decision and Control*. 2005.

[2] J. P. Cunningham, K. V. Shenoy, and M. Sahani, "Fast Gaussian process methods for point process intensity estimation," in *Proc. of the 25th Ann. Int. Conf. on Machine Learning (ICML 2008)*, Andrew McCallum and Sam Roweis, Eds., pp. 192–199. 2008.

[3] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME, Journal of Basic Engineering*, vol. 82, pp. 34–45, March 1960.

[4] N. Wiener, *Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications*, John Wiley & Sons, Inc., New York, 1950.

[5] H. L. Van Trees, *Detection, Estimation, and Modulation Theory Part I*, John Wiley & Sons, New York, 1968.

[6] H. L. Van Trees, *Detection, Estimation, and Modulation Theory Part II*, John Wiley & Sons, New York, 1971.

[7] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.

[8] R. L. Stratonovich, *Topics in the Theory of Random Noise*, Gordon and Breach, 1963.

[9] R. E. Kalman and R. S. Bucy, "New results in linear filtering and prediction theory," *Transactions of the ASME, Journal of Basic Engineering*, vol. 83, pp. 95–108, March 1961.

[10] H. E. Rauch, F. Tung, and C. T. Striebel, "Maximum likelihood estimates of linear dynamic systems," *AIAA Journal*, vol. 3(8), pp. 1445–1450, 1965.

[11] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*, Academic Press, 1970.

[12] A. O'Hagan, "Curve fitting and optimal design for prediction (with discussion)," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 40(1), pp. 1–42, 1978.

[13] M. S. Grewal and A. P. Andrews, *Kalman Filtering, Theory and Practice Using MATLAB*, Wiley Interscience, 2001.

[14] T. Glad and L. Ljung, *Control Theory: Multivariable and Nonlinear Methods*, Taylor & Francis, 2000.

[15] Y. Bar-Shalom, X. Li, and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation*, Wiley Interscience, 2001.

[16] P. Whittle, "On stationary processes in the plane," *Biometrica*, vol. 41(3/4), pp. 434–449, 1954.

[17] B. Matérn, "Spatial variation – stochastic models and their application to some problems in forest surveys and other sampling investigations," Tech. Rep., Meddelanden från Statens Skogforskningsinstitut, 1960, Band 49 - Nr 5.