Spatio-Temporal Learning via Infinite-Dimensional Bayesian Filtering and Smoothing

Simo Särkkä, *Senior Member, IEEE,* Arno Solin, and Jouni Hartikainen

***Abstract*—Gaussian process based machine learning is a powerful Bayesian paradigm for non-parametric non-linear regression and classification. In this paper, we discuss connections of Gaussian process regression with Kalman filtering, and present methods for converting spatio-temporal Gaussian process regression and classification problems into infinite-dimensional state space models. This formulation allows for use of computationally efficient infinite-dimensional Kalman filtering and smoothing methods, or more general Bayesian filtering and smoothing methods, which reduces the problematic cubic complexity of Gaussian process regression in the number of time steps into linear time complexity. The implication of this is that the use of machine learning models in signal processing becomes computationally feasible, and it opens the possibility to combine machine learning techniques with signal processing methods.**

***Index Terms*—Gaussian process, Gaussian random field, infinite-dimensional state space model, infinite-dimensional Kalman filtering and smoothing**

I. Introduction

SPATIO-temporal Gaussian processes, or Gaussian fields, arise in many disciplines such as spatial statistics and kriging, machine learning, physical inverse problems, and signal processing [1], [2], [3], [4], [5]. In these applications, we are interested in doing statistical inference on the dynamic state of the whole field based on a finite set of indirect measurements as well as estimating the properties (i.e., the parameters) of the underlying process (or field). For example, in electrical impedance tomography (EIT) problems [4] we try to reconstruct the resistance field of a body based on voltages induced by injected currents. In spatial statistics typical problems are prediction of wind, precipitation or ocean currents based on finite sets of measurements [1].

In Gaussian process based Bayesian machine learning [2] Gaussian processes are used as non-parametric priors for regressor functions, and the space and time variables take the roles of input variables of the regressor function. Learning in these non-parametric models amounts to computation of the posterior distribution of the Gaussian process conditioned on a set of measurements and estimation of the parameters of the covariance function of the process. One way to interpret Gaussian process regression is to see it as a kernel method, where the covariance function of the Gaussian process serves as the kernel of the corresponding reproducing kernel Hilbert space (RKHS) (cf. [6]).

S. Särkkä and Arno Solin are with the Department of Biomedical Engineering and Computational Science (BECS), Aalto University, 02150 Espoo, Finland e-mail: { simo.sarkka, arno.solin }@aalto.fi.

J. Hartikainen is with Anonymous University.

Manuscript received January 8, 2013; revised January 8, 2013.

In classical signal processing and stochastic control, Gaussian processes are commonly used for modeling temporal phenomena in form of stochastic differential equations (SDE) [7] and the inference procedure is usually solved using Kalman filter type of methods [5]. These models are typically based on physical, electrical or mechanical principles, which can be represented in form of differential equations. Stochasticity in these systems is in form of a white noise process acting as an unknown forcing term. It is also possible to use more general Gaussian process models to better model the forcing terms in such systems. This is the underlying idea in latent force models (LFM), which have many applications, for example, in biology, human tracking and satellite navigation [8], [9].

A central practical problem in the Gaussian process regression context as well as in more general statistical inverse problems is the cubic $O(N^3)$ computational complexity in the number of measurements N. In the spatio-temporal setting, when we obtain, say, M measurements per time step and the total number of time steps is T, this translates into a cubic complexity in space and time $O(M^3\,T^3)$. This is problematic in signal processing, because there we usually are interested in processing long (unbounded) time series and thus it is necessary to have linear complexity in time. This is also the reason for the popularity of SDE models and state space models in signal processing—their inference problem can be solved with Kalman (or Bayesian) filters and smoothers which have a linear $O(T)$ time complexity.

Due to the computational efficiency of Kalman filters and smoothers, it is beneficial to reformulate certain spatio-temporal Gaussian process regression problems as Kalman filtering and smoothing problems. The aim of this paper is to show when and how this is possible. We also present a number of analytical examples, and apply the methodology to prediction of precipitation and to processing of fMRI brain imaging data.

The described methods will be mainly based on the articles Hartikainen & Särkkä [10] and Särkkä & Hartikainen [11]. However, the idea of reducing the computational complexity of Gaussian process regression (or equivalent kriging) via reduction into SDE form has also been suggested by Lindgren et al. [12], and filtering and smoothing type of methods have been applied to spatio-temporal context before [13], [4], [14]. Applying recursive Bayesian methods to on-line learning in Gaussian process regression has also been proposed, for example, in [15] and they are also closely related to kernel recursive least squares (KRLS) algorithms [16], [17]. Infinite-dimensional filtering and smoothing methods as such date back to the 60's–70's [18].

The structure of this paper is as follows. In Section II we

describe how Gaussian processes are used in regression and Kalman filtering, and what the idea behind combining the approaches is. In Section III we present methods for converting Gaussian process regression models into state space models which are suitable for Kalman filtering and smoothing methods. In Section IV we discuss how the inference procedure can be done in practice, how it can be extended to non-linear and non-Gaussian models and parameter estimation, and finally in Section V, we present two example applications.

II. Gaussian Processes in Regression and Kalman Filtering

A. Definition of a Gaussian Process

A Gaussian process is a random function $f(\boldsymbol{\xi})$ with d-dimensional input $\boldsymbol{\xi}$ such that any finite collection of random variables $f(\boldsymbol{\xi}_1), \ldots, f(\boldsymbol{\xi}_n)$ has a multidimensional Gaussian distribution. Note that when $d > 1$, what we here call Gaussian processes are often called Gaussian fields, but here we will always use the term *process*, regardless of the dimensionality d.

A Gaussian process can be defined in terms of a mean $m(\boldsymbol{\xi})$ and covariance function $k(\boldsymbol{\xi}, \boldsymbol{\xi}')$:

$$\begin{aligned} m(\boldsymbol{\xi}) &= \mathrm{E}[f(\boldsymbol{\xi})] \\ k(\boldsymbol{\xi}, \boldsymbol{\xi}') &= \mathrm{E}[(f(\boldsymbol{\xi}) - m(\boldsymbol{\xi}))\,(f(\boldsymbol{\xi}') - m(\boldsymbol{\xi}'))]. \end{aligned} \tag{1}$$

The joint distribution of an arbitrary finite collection of random variables $f(\boldsymbol{\xi}_1), \ldots, f(\boldsymbol{\xi}_n)$ is then multidimensional Gaussian:

$$\begin{pmatrix} f(\boldsymbol{\xi}_1) \\ \vdots \\ f(\boldsymbol{\xi}_n) \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} m(\boldsymbol{\xi}_1) \\ \vdots \\ m(\boldsymbol{\xi}_n) \end{pmatrix}, \begin{pmatrix} k(\boldsymbol{\xi}_1, \boldsymbol{\xi}_1) & \ldots & k(\boldsymbol{\xi}_1, \boldsymbol{\xi}_n) \\ \vdots & \ddots & \\ k(\boldsymbol{\xi}_n, \boldsymbol{\xi}_1) & & k(\boldsymbol{\xi}_n, \boldsymbol{\xi}_n) \end{pmatrix} \right) \tag{2}$$

In the same way as a function can be considered as an infinite-long vector consisting of all its values at each input, one way to think about a Gaussian process is to consider it as an infinite-dimensional limit of a Gaussian random vector. The input variable then serves as the element index in this infinite-long random vector.

A Gaussian process is *stationary* if its mean is constant and the two-argument covariance function is of the form

$$k(\boldsymbol{\xi}, \boldsymbol{\xi}') = C(\boldsymbol{\xi}' - \boldsymbol{\xi}). \tag{3}$$

where $C(\boldsymbol{\xi})$ is another function, the stationary covariance function of the process.

In spatial Gaussian processes, we have $\boldsymbol{\xi} = \mathbf{x}$, where $\mathbf{x}$ corresponds to the input of the random function in Gaussian process regression or, for example, a spatial location in geostatistics or physical inverse problems. In temporal Gaussian processes typically arising in signal processing, $\boldsymbol{\xi} = t$, where t is the time variable. In spatio-temporal problems $\boldsymbol{\xi} = (\mathbf{x}, t)$, where $\mathbf{x}$ and t are the space (or input) and time variables, respectively.

B. Gaussian Process Regression

Gaussian process regression is a way to do non-parametric regression with Gaussian processes. Assume that we want to predict (interpolate) the values of an unknown scalar valued function with d-dimensional input:

$$y = f(\mathbf{x}) \tag{4}$$

at a certain test point $\mathbf{x}^*$, based on a training set consisting of a finite number of observed input-output pairs $\{(\mathbf{x}_i, y_i) : i = 1, \ldots, N\}$. Instead of postulating a parametric form of the function $f(\mathbf{x}, \boldsymbol{\theta})$ as in parametric regression and estimating the parameters $\boldsymbol{\theta}$, in Gaussian process regression, we assume that the function $f(\mathbf{x})$ is a sample from Gaussian process with a given mean $m(\mathbf{x})$ and covariance function $k(\mathbf{x}, \mathbf{x}')$. This is denoted as follows:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')). \tag{5}$$

In Gaussian process regression, we typically use stationary covariance functions and assume that the mean is identically zero $m(\mathbf{x}) = 0$. A very common choice of covariance function is the squared exponential covariance function

$$k(\mathbf{x}, \mathbf{x}') = s^2 \exp\left(-\frac{1}{2l^2} ||\mathbf{x} - \mathbf{x}'||^2 \right), \tag{6}$$

which has the property that the resulting sample functions are very smooth (infinitely differentiable). The parameters l and s then define how smooth the function actually is and what is the magnitude of values that we should expect.

The underlying idea in Gaussian process regression is that the correlation structure introduces dependencies between function values at different inputs. Thus the function values at the observed points give information also of the unobserved points. For example, the squared exponential covariance function above says that when the inputs are close to each other, also the function values should be close to each other. This is equivalent to saying that the function values with similar inputs should have a stronger correlation than function values with dissimilar inputs, which is exactly what the above covariance function states.

In statistical estimation problems it is often assumed that the measurements are not perfect, but instead, they are corrupted by certain additive Gaussian noise. That is, the measurements are modeled as

$$y_i = f(\mathbf{x}_i) + \varepsilon_i, \qquad \varepsilon_i \sim \mathrm{N}(0, \sigma^2), \tag{7}$$

where ε_i are IID random variables, a priori independent of the Gaussian process $f(\mathbf{x})$. Now, we are interested in computing an estimate of the value of the "clean" function $f(\mathbf{x}^*)$ at test point $\mathbf{x}^*$.

The key thing is now to observe that the joint distribution of the test point and the training points $(f(\mathbf{x}^*), f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N))$ is Gaussian with known statistics (this follows from the definition of a Gaussian process). Because the measurement model (7) is linear-Gaussian, the joint distribution of the test point and the measurements $(f(\mathbf{x}^*), y_1, \ldots, y_N)$ is Gaussian with know statistics as well. As everything is Gaussian, we can compute the conditional distribution of $f(\mathbf{x}^*)$ given the observations $y_1, \ldots, y_N$ by using the well-known computational rules for Gaussian distributions. The result can be expressed as

$$p(f(\mathbf{x}^*) \,|\, y_1, \ldots, y_N) = \mathrm{N}(f(\mathbf{x}^*) \,|\, \mu(\mathbf{x}^*), V(\mathbf{x}^*)), \tag{8}$$

where the posterior mean $\mu(\mathbf{x}^*)$ is a function of the training inputs and the measurements, and the posterior variance $V(\mathbf{x}^*)$ is a function of the training inputs (see, [2] for details). However, it turns out that the computational complexity of the equations is $O(N^3)$, because of an $N \times N$-matrix inversion appearing in both the mean and covariance equations.

A useful way to rewrite the Gaussian process regression problem is in the form

$$\begin{aligned} f(\mathbf{x}) &\sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \\ \mathbf{y} &= \mathcal{H} f(\mathbf{x}) + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \Sigma), \end{aligned} \tag{9}$$

where $\Sigma = \sigma^2 \mathbf{I}$, and the linear operator $\mathcal{H}$ picks the training set inputs among the function values:

$$\mathcal{H} f(\mathbf{x}) = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N)). \tag{10}$$

This problem can be seen as a infinite-dimensional version of the following Bayesian linear regression problem:

$$\begin{aligned} \mathbf{f} &\sim \mathrm{N}(\mathbf{m}, \mathbf{K}) \\ \mathbf{y} &= \mathbf{H}\,\mathbf{f} + \boldsymbol{\varepsilon}, \end{aligned} \tag{11}$$

where $\mathbf{f}$ is a vector with Gaussian prior $\mathrm{N}(\mathbf{m}, \mathbf{K})$, and matrix $\mathbf{H}$ is constructed to pick those elements of the vector $\mathbf{f}$ that we actually observe. Solving the infinite-dimensional linear regression problem in Equation (9) analogously to the finite-dimensional problem in Equation (11), leads to the Gaussian process regression solution (cf. [11]). We could also replace the operator $\mathcal{H}$ with a more general linear operator which would allow us to solve statistical inverse problems under the Gaussian process regression formalism [19].

C. Kalman Filtering and Smoothing

Kalman filtering is considered with statistical inference on state space models of the form

$$\begin{aligned} \frac{\mathrm{d}\mathbf{f}(t)}{\mathrm{d}t} &= \mathbf{A}\,\mathbf{f}(t) + \mathbf{L}\,\mathbf{w}(t) \\ \mathbf{y}_k &= \mathbf{H}\,\mathbf{f}(t_k) + \boldsymbol{\varepsilon}_k, \end{aligned} \tag{12}$$

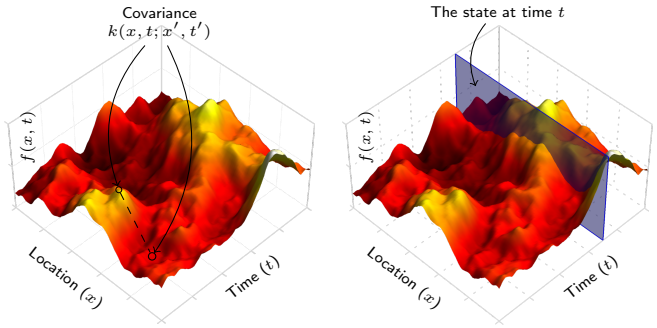
where $k = 1, \ldots, T$, and $\mathbf{A}$, $\mathbf{L}$, and $\mathbf{H}$ are given matrices, $\boldsymbol{\varepsilon}_k$ is a vector of Gaussian measurement noises, and $\mathbf{w}(t)$ is a vector of white noise processes. A *white noise process* refers to a zero mean Gaussian random process, where each pair of values $\mathbf{w}(t)$ and $\mathbf{w}(t')$ are uncorrelated whenever $t \neq t'$.

Because $\mathbf{f}(t)$ is a solution to a linear differential equation driven by Gaussian noise, it is a Gaussian process. It is also possible to compute the corresponding covariance function of $\mathbf{f}(t)$ (see, e.g., [10]), which gives the equivalent covariance function based formulation. We can also construct almost any covariance function for a single selected component of the state provided that we augment the state to contain a number of derivatives of the selected state component as well [10].

Because the solution of a stochastic differential equation is a Markovian process, it allows for linear time computation of the posterior distribution of any unobserved test point $\mathbf{f}(t^*)$. The computational algorithms for this are the Kalman filter and Rauch–Tung–Striebel (RTS) smoother algorithm. The Kalman filter and RTS smoother can be used for computing the mean and covariance of the following distribution for arbitrary t in linear time complexity:

$$p(\mathbf{f}(t) \,|\, \mathbf{y}_1, \ldots, \mathbf{y}_T) = \mathrm{N}(\mathbf{f}(t) \,|\, \mathbf{m}_s(t), \mathbf{P}_s(t)). \tag{13}$$

Thus we can now pick the time point $t = t^*$ to obtain the posterior distribution of $\mathbf{f}(t^*)$.



(a) Covariance function view (b) State space model view

Fig. 1. **(a)** In the covariance function based representation the spatio-temporal field is considered "frozen" and we postulate the covariance between two space–time points. **(b)** In the state space model based description we construct a differential equation for the temporal behavior of a sequence of "snapshots" of the spatial field.

D. Combining the Approaches

Spatio-temporal Gaussian process regression is considered with models of the form

$$\begin{aligned} f(\mathbf{x}, t) &\sim \mathcal{GP}(0, k(\mathbf{x}, t; \mathbf{x}', t')) \\ \mathbf{y}_k &= \mathcal{H}_k f(\mathbf{x}, t_k) + \boldsymbol{\varepsilon}_k, \end{aligned} \tag{14}$$

which we have already written in form similar to Equation (9).

As mentioned in the previous section, it is possible to force almost any covariance function of a given state component of a state space model provided that we augment the state with a sufficient number of time derivatives as well. We can now use this idea to formulate a hybrid model, where the temporal correlation in the above model is represented as a stochastic differential equation kind of model and the spatial correlation is injected into the model by selecting the matrices in the model properly. In fact, we need to let the spatial dimension to take the role of an additional vector element index, which leads to the infinite-dimensional state space model

$$\begin{aligned} \frac{\partial \mathbf{f}(\mathbf{x}, t)}{\partial t} &= \mathcal{A}\,\mathbf{f}(\mathbf{x}, t) + \mathbf{L}\,\mathbf{w}(\mathbf{x}, t) \\ \mathbf{y}_k &= \mathcal{H}_k\,\mathbf{f}(\mathbf{x}, t_k) + \boldsymbol{\varepsilon}_k, \end{aligned} \tag{15}$$

where the state $\mathbf{f}(\mathbf{x}, t)$ at time t consists of the whole function $\mathbf{x} \mapsto f(\mathbf{x}, t)$ and a suitable number its time derivatives. This model is now an infinite-dimensional Markovian type of model which allows for linear-time inference with the infinite-dimensional Kalman filter and RTS smoother.

The philosophical difference between the covariance function based model in Equation (14) and the state space model in Equation (15) is illustrated in Figure 1. In the state space model formulation we think that we have a field which propagates forward in space whereas in the covariance based

formulation we just compute covariances between space–time points of a "frozen" field.

III. Converting Covariance Functions to State Space Models

A. Covariance Functions of Stochastic Differential Equations

One useful way to construct Gaussian processes is as solutions to nth order stochastic linear differential equations of the form

$$a_n \frac{\mathrm{d}^n f(t)}{\mathrm{d}t^n} + \cdots + a_1 \frac{\mathrm{d}f(t)}{\mathrm{d}t} + a_0 f(t) = w(t), \qquad (16)$$

where the driving function $w(t)$ is a zero-mean continuous-time Gaussian white noise process. The solution process $f(t)$, a random function, is a Gaussian process, because $w(t)$ is Gaussian and the solution of a linear differential equation is a linear operation on the input.

If we take the formal Fourier transform of Equation (16) and solve for the Fourier transform of the process $F(i\,\omega)$, we get

$$F(i\omega) = \underbrace{\left(\frac{1}{a_n\,(i\,\omega)^n + \cdots + a_1\,(i\,\omega) + a_0}\right)}_{G(i\,\omega)} W(i\,\omega), \quad (17)$$

where $W(i\,\omega)$ is the (formal) Fourier transform of the white noise. The above equation can be interpreted such that the process $F(i\,\omega)$ is obtained by feeding white noise though a system with the transfer function $G(i\,\omega)$.

From the above description it is now easy to calculate the corresponding (power) spectral density of the process, which is just the square of the absolute value of the Fourier transform of the process. If we denote the spectral density of white noise $|W(i\,\omega)|^2 = q_c$, the spectral density of the process is

$$S(\omega) = q_c\,|G(i\,\omega)|^2, \qquad (18)$$

where the key factor is to observe that it has the form

$$S(\omega) = \frac{\text{constant}}{\text{polynomial in } \omega^2}. \qquad (19)$$

Thus we can conclude that for an nth order random differential equation of the form (16) the spectral density has the rational function form.

The classical Wiener–Khinchin theorem states that the stationary covariance function of the process is given by the inverse Fourier transform of the spectral density:

$$C(t) = \mathcal{F}^{-1}[S(\omega)] = \frac{1}{2\pi}\int S(\omega)\,\exp(i\,\omega\,t)\,\mathrm{d}\omega, \qquad (20)$$

and thus the corresponding covariance function is

$$k(t, t') = C(t' - t). \qquad (21)$$

Note that the stochastic differential equation (16) can also be equivalently represented in the following state space form. If we define $\mathbf{f} = (f, \mathrm{d}f/\mathrm{d}t, \ldots, \mathrm{d}^{n-1}f/\mathrm{d}t^{n-1})$, then we have

$$\frac{\mathrm{d}\mathbf{f}(t)}{\mathrm{d}t} = \underbrace{\begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ -a_0 & -a_1 & \ldots & -a_{n-1} \end{pmatrix}}_{\mathbf{A}} \mathbf{f}(t) + \underbrace{\begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}}_{\mathbf{L}} w(t). \qquad (22)$$

Recall that the scalar function $f(t)$ is just the first component of the vector $\mathbf{f}(t)$. Thus if we assume that we measure noise corrupted values y_k of $f(t_k)$ at points $t_1, \ldots, t_N$, we can write this as

$$y_k = \underbrace{\begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix}}_{\mathbf{H}} \mathbf{f}(t) + \varepsilon_k, \qquad (23)$$

which indeed is a model of the form (12).

Thus if we apply the Kalman filter and smoother to the state space model described by the Equations (22) and (23) we will get the same result as if we applied Gaussian process regression equations to the covariance function defined by the Equation (20). In that sense the representations are equivalent. However, if the number of time steps is T, then the computational complexity of the Kalman filter and smoother is $O(T)$, whereas the complexity of Gaussian process regression solution is $O(T^3)$. Thus the state space formulation has a huge computational advantage, at least when the number of time steps is large.

In fact, we do not need to restrict ourselves to spectral densities of the all-pole form (19), but the transfer function in Equation (17) can be allowed to have the more general form

$$G(i\,\omega) = \frac{b_m\,(i\,\omega)^m + \cdots + b_1\,(i\,\omega) + b_0}{a_n\,(i\,\omega)^n + \cdots + a_1\,(i\,\omega) + a_0}, \qquad (24)$$

where the numerator order is strictly lower than the denominator order $m < n$ with $a_n \neq 0$ (i.e., the transfer function is proper). The spectral density then has the more general form

$$S(\omega) = \frac{m\text{th order polynomial in } \omega^2}{n\text{th order polynomial in } \omega^2}, \qquad (25)$$

where the order of the numerator is again lower than the denominator's.

In control theory [20], there exists a number of methods to convert a transfer function of form (24) into an equivalent state space model. The procedure which we used above roughly corresponds to the so called *controller canonical form* of the state space model. Another option is, for example, the *observer canonical form*. In these representations the state variables are not necessarily pure time derivatives anymore. Anyway, once we have the transfer function, we can always convert it into a state space model.

B. From Temporal Covariance Functions to State Space Models

An interesting question is now that if we are given a covariance function $k(t, t')$, how can we find a state space model where one of the state components has this covariance

function? We will assume that the process is stationary and thus we can equivalently say that we want to find a state space model with a component having a stationary covariance function $C(t)$ such that $k(t,t') = C(t-t')$.

The procedure to *convert a covariance function into state space model* is the following [10]:

1) Compute the corresponding spectral density $S(\omega)$ by computing the Fourier transform of $C(t)$.
2) If $S(\omega)$ is not a rational function of the form (25), then approximate it with such a function. This approximation can be formed using, for example, Taylor series expansions or Padé approximants.
3) Find a *stable* rational transfer function $G(i\,\omega)$ of the form (24) and constant q_c such that
$$S(\omega) = G(i\,\omega)\,q_c\,G(-i\,\omega). \tag{26}$$
The procedure for finding a stable transfer function is called spectral factorization. One method to do that is outlined later in this section.
4) Use the methods from control theory [20] to convert the transfer function model into an equivalent state space model. The constant q_c will then be the spectral density of the driving white noise process.

An example of the above procedure is presented in Example 1 for the Matérn covariance function. However, above we required the transfer function $G(i\,\omega)$ to be *stable*. A transfer function corresponds to a stable system if and only if all of its poles (zeros of the denominator) are in the upper half of the complex plane. We also want the transfer function to be minimum phase, which happens when the zeros of the numerator are also in the upper half plane.

The procedure to find such a transfer function is called *spectral factorization*. One simple way to do that is the following:

- Compute the roots of the numerator and denominator polynomials of $S(\omega)$. The roots will appear in pairs, where one member of the pair is always the complex conjugate of the other.
- Construct the numerator and denominator polynomials of the transfer function $G(i\,\omega)$ from the positive-imaginary-part roots only.

If the spectral density does not already have a rational function form, the above procedures only lead to approximations. One example of a covariance function which does not have a rational spectral density is the squared exponential covariance function in Equation (6). However, its spectral density can be well approximated with low order rational functions. This is demonstrated in Example 2.

Example 1 (1D Matérn covariance function)**.** *The isotropic and stationary ($\tau = |t-t'|$) covariance function of the Matérn family can be given as*
$$C(\tau) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\,\frac{\tau}{l}\right)^{\nu} K_\nu\left(\sqrt{2\nu}\,\frac{\tau}{l}\right),$$
where $\nu, \sigma, l > 0$ are the smoothness, magnitude and length scale parameters, and $K_\nu(\cdot)$ the modified Bessel function (see, e.g., [2]). The spectral density is of the form
$$S(\omega) \propto \left(\lambda^2 + \omega^2\right)^{-(\nu+1/2)},$$
where $\lambda = \sqrt{2\nu}/l$. Thus the spectral density can be factored as $S(\omega) \propto (\lambda + i\,\omega)^{-(p+1)}\,(\lambda - i\,\omega)^{-(p+1)}$, where $\nu = p + 1/2$. The transfer function of the corresponding stable part is
$$G(i\,\omega) = (\lambda + i\,\omega)^{-(p+1)}.$$
For integer values of p ($\nu = 1/2, 3/2, \ldots$), we can expand this expression using the binomial formula. For example, if $p = 1$ ($\nu = 3/2$), the corresponding LTI SDE is
$$\frac{\mathrm{d}\mathbf{f}(t)}{\mathrm{d}t} = \begin{pmatrix} 0 & 1 \\ -\lambda^2 & -2\lambda \end{pmatrix} \mathbf{f}(t) + \begin{pmatrix} 0 \\ 1 \end{pmatrix} w(t),$$
where $\mathbf{f}(t) = (f(t), \mathrm{d}f(t)/\mathrm{d}t)$. The covariance, spectral density and an example realization are shown in Figure 2.

Example 2 (1D squared exponential covariance function)**.** *The one-dimensional squared exponential covariance function*
$$C(\tau) = \sigma^2 \exp(-\tau^2/(2l^2))$$
has the spectral density
$$S(\omega) = \sigma^2 \sqrt{2\pi}\, l\, \exp\left(-\frac{l^2\,\omega^2}{2}\right).$$
This spectral density is not a rational function, but we can easily approximate it with such a function. By using the Taylor series of the exponential function we get an approximation
$$S(\omega) \approx \frac{\textit{constant}}{1 + l^2\,\omega^2 + \cdots + \frac{1}{n!}\,l^{2n}\,\omega^{2n}}$$
which we can factor into stable and unstable parts, and further convert into an n-dimensional state space model
$$\frac{\mathrm{d}\mathbf{f}(t)}{\mathrm{d}t} = \mathbf{A}\,\mathbf{f}(t) + \mathbf{L}\,w(t).$$
The covariance, spectral density and a random realization are shown in Figure 2. The error induced by the Taylor series expansion approximation is also illustrated in the figure.

C. State Space Representation of Spatio-Temporal Gaussian Processes

Let's now consider the question of representing a spatio-temporal covariance function $k(\mathbf{x}, t; \mathbf{x}', t')$ in state space form. Assuming that the covariance function is stationary we can concentrate on the corresponding stationary covariance function $C(\mathbf{x}, t)$. The space–time Fourier transform then gives the corresponding spectral density $S(\boldsymbol{\omega}_x, \omega_t)$.

If we consider $\boldsymbol{\omega}_x$ fixed, then $\omega_t \mapsto S(\boldsymbol{\omega}_x, \omega_t)$ can be considered as a spectral density of a temporal process which is parametrized with $\boldsymbol{\omega}_x$. Assume for simplicity that the function $\omega_t \mapsto S(\boldsymbol{\omega}_x, \omega_t)$ has the form of a constant divided by polynomials (19). This implies that there exists an nth order spatial Fourier domain stochastic differential equation which

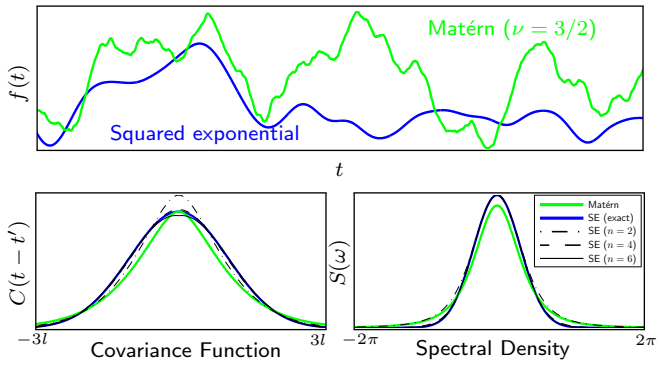


Fig. 2. Random realizations drawn using the state space models in Examples 1 (green) and 2 (blue). The processes can characterized through their covariance functions or using their spectral densities. The representation for the Matérn covariance function is exact whereas the squared exponential needs to be approximated with a finite-order model (illustrated with $n = 2, 4, 6$ above). The errors in the tails of the spectral density transform into bias at the origin of the covariance function. With order $n = 6$, which was also used for drawing the random realization, both the approximations are already almost indistinguishable from the exact ones.

has the same temporal spectral density:

$$a_n(i\,\boldsymbol{\omega}_x)\,\frac{\partial^n \tilde{f}(i\,\boldsymbol{\omega}_x,t)}{\partial t^n} + \cdots + a_1(i\,\boldsymbol{\omega}_x)\,\frac{\partial \tilde{f}(i\,\boldsymbol{\omega}_x,t)}{\partial t} + a_0(i\,\boldsymbol{\omega}_x)\,\tilde{f}(i\,\boldsymbol{\omega}_x,t) = \tilde{w}(i\,\boldsymbol{\omega}_x,t), \tag{27}$$

where the spectral density of the white noise process is some function $\tilde{q}_c(\boldsymbol{\omega}_x)$. Analogously to the temporal case (cf. Section III-A) we can now express this in the following equivalent state space form:

$$\frac{\partial \tilde{\mathbf{f}}(i\,\boldsymbol{\omega}_x,t)}{\partial t} = \mathbf{A}(i\,\omega_x)\,\tilde{\mathbf{f}}(i\,\boldsymbol{\omega}_x,t) + \mathbf{L}\,\tilde{w}(i\,\boldsymbol{\omega}_x,t), \tag{28}$$

where

$$\mathbf{A}(i\,\omega_x) = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ -a_0(i\,\boldsymbol{\omega}_x) & -a_1(i\,\boldsymbol{\omega}_x) & \ldots & -a_{n-1}(i\,\boldsymbol{\omega}_x) \end{pmatrix}, \tag{29}$$

and $\mathbf{L}$ is the same as in Equation (22).

The above equation is still in spatial Fourier domain and to convert it into spatial domain, we need to compute its inverse Fourier transform. This leads to the equation

$$\frac{\partial \mathbf{f}(\mathbf{x},t)}{\partial t} = \boldsymbol{\mathcal{A}}\,\mathbf{f}(\mathbf{x},t) + \mathbf{L}\,w(\mathbf{x},t), \tag{30}$$

where $\boldsymbol{\mathcal{A}}$ is a matrix of linear operators as follows:

$$\boldsymbol{\mathcal{A}} = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ -\mathcal{A}_0 & -\mathcal{A}_1 & \ldots & -\mathcal{A}_{n-1} \end{pmatrix}. \tag{31}$$

and the spatial covariance of the white noise is given by the inverse Fourier transform of $\tilde{q}_c(\boldsymbol{\omega}_x)$. The operators $\mathcal{A}_j$ are pseudo-differential operators [21] defined in terms of their Fourier transforms:

$$\begin{aligned} \mathcal{A}_0 &= \mathcal{F}_x^{-1}[a_0(i\,\boldsymbol{\omega}_x)], \\ \mathcal{A}_1 &= \mathcal{F}_x^{-1}[a_1(i\,\boldsymbol{\omega}_x)], \\ &\vdots \\ \mathcal{A}_{n-1} &= \mathcal{F}_x^{-1}[a_{n-1}(i\,\boldsymbol{\omega}_x)]. \end{aligned} \tag{32}$$

The measurement model operator $\mathcal{H}$ can now be constructed such that it evaluates the first component of the vector $\mathbf{f}(\mathbf{x},t)$ at the measurement points by combining the ideas in Equations (10) and (23).

Analogously to the temporal case, we can generalize the above procedure to models with transfer functions of the form

$$G(i\,\omega_t) = \frac{b_m(i\boldsymbol{\omega}_x)\,(i\,\omega_t)^m + \cdots + b_1(i\boldsymbol{\omega}_x)\,(i\,\omega_t) + b_0(i\boldsymbol{\omega}_x)}{a_n(i\boldsymbol{\omega}_x)\,(i\,\omega_t)^n + \cdots + a_1(i\boldsymbol{\omega}_x)\,(i\,\omega_t) + a_0(i\boldsymbol{\omega}_x)}, \tag{33}$$

which then leads to spectral densities $S(\boldsymbol{\omega}_x,\omega_t)$ which are rational in variable ω_t^2.

D. From Spatio-Temporal Covariance Functions to State Space Models

The generalization of the conversion procedure presented in Section III-B is the following. A given stationary spatio-temporal Gaussian process with covariance function $k(\mathbf{x},t;\mathbf{x}',t')$ such that $k(\mathbf{x},t;\mathbf{x}',t') = C(\mathbf{x}'-\mathbf{x},t'-t)$ can be converted into an infinite-dimensional state space model representation via the following steps:

1) Compute the corresponding spectral density $S(\boldsymbol{\omega}_x,\omega_t)$ as the spatio-temporal Fourier transform of $C(\mathbf{x},t)$.
2) Approximate the function $\omega_t \mapsto S(\boldsymbol{\omega}_x,\omega_t)$ with a rational function in variable ω_t^2.
3) Find a stable ω_t-rational transfer function $G(i\,\boldsymbol{\omega}_x, i\,\omega_t)$ and function $\tilde{q}_c(\boldsymbol{\omega}_x)$ such that
$$S(\boldsymbol{\omega}_x,\omega_t) = G(i\,\boldsymbol{\omega}_x, i\,\omega_t)\,\tilde{q}_c(\boldsymbol{\omega}_x)\,G(-i\,\boldsymbol{\omega}_x, -i\,\omega_t). \tag{34}$$
The transfer function needs to have all its roots and zeros with respect to the ω_t variable in upper half plane, for all values of $\boldsymbol{\omega}_x$. This kind of representation can be found using spectral factorization discussed in Section III-B.
4) Use the methods from control theory [20] to convert the transfer function model into an equivalent spatial Fourier domain state space model.
5) Transform each of the coefficients $a_j(i\,\boldsymbol{\omega}_x)$ and $b_j(i\,\boldsymbol{\omega}_x)$ into the corresponding pseudo-differential operators and set the spatial stationary covariance function of the white noise process to the inverse Fourier transform of $\tilde{q}_c(\boldsymbol{\omega}_x)$.

The above procedure is demonstrated in Example 3 for the spatio-temporal Matérn covariance function.

Note that when the covariance function is separable, that is, $C(\mathbf{x},t) = C_x(\mathbf{x})\,C_t(t)$, it implies that the spectral density is separable as well: $S(\boldsymbol{\omega}_x,\omega_t) = S_x(\boldsymbol{\omega}_x)\,S_t(\omega_t)$. It now turns out that we can do the factorization in Equation (34) as follows:

$$S(\boldsymbol{\omega}_x,\omega_t) = G(i\,\omega_t)\,S_x(\boldsymbol{\omega}_x)\,G(-i\,\omega_t). \tag{35}$$

Example 3 (2D Matérn covariance function)**.** *The multi-dimensional equivalent of the Matérn covariance function given in Example 1 is the following ($r = \|\boldsymbol{\xi} - \boldsymbol{\xi}'\|$, for $\boldsymbol{\xi} = (x_1, x_2, \ldots, x_{d-1}, t) \in \mathbb{R}^d$):*

$$C(r) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\,\frac{r}{l}\right)^{\nu} K_\nu \left(\sqrt{2\nu}\,\frac{r}{l}\right).$$

The corresponding spectral density is of the form

$$S(\boldsymbol{\omega}_r) = S(\boldsymbol{\omega}_x, \omega_t) \propto \frac{1}{(\lambda^2 + \|\boldsymbol{\omega}_x\|^2 + \omega_t^2)^{\nu+d/2}}.$$

where $\lambda = \sqrt{2\nu}/l$. In order to find the transfer function $G(i\boldsymbol{\omega}_x, i\omega_t)$, we find the roots of the expression in the denominator. They are given by $(i\omega_t) = \pm\sqrt{\lambda^2 - (i\boldsymbol{\omega}_x)^2}$, which means we can now extract the transfer function of the stable Markov process

$$G(i\boldsymbol{\omega}_x, i\omega_t) = \left(i\omega_t + \sqrt{\lambda^2 - \|i\boldsymbol{\omega}_x\|^2}\right)^{-(\nu+d/2)}$$

The expansion of the denominator depends on the value of $p = \nu + d/2$. If p is an integer, the expansion can be easily done by the binomial theorem. For example, if $\nu = 1$ and $d = 2$, we get the following

$$\frac{\partial \mathbf{f}(x,t)}{\partial t} = \begin{pmatrix} 0 & 1 \\ -(\lambda^2 - \nabla^2) & -2\sqrt{\lambda^2 - \nabla^2} \end{pmatrix} \mathbf{f}(x,t) + \begin{pmatrix} 0 \\ 1 \end{pmatrix} w(x,t), \tag{36}$$

where ∇^2 is the (spatial) Laplace operator (here the second partial derivative w.r.t. x). The one-dimensional example in Example 1 can be seen as a special case of this. An example realization of the process is shown in Figure 3.

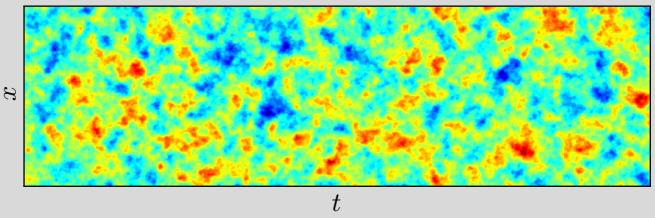


Fig. 3: A random realization simulated by the state space model in Equation (36)*.*

Because the transfer function $G(i\,\omega_t)$ does not contain the variable $\boldsymbol{\omega}_x$ at all, the operator matrix $\boldsymbol{\mathcal{A}}$ will actually be just an ordinary matrix and the space correlation gets accounted by setting the spatial covariance of the white noise process according to the spectral density $S_x(\boldsymbol{\omega}_x)$. The resulting infinite-dimensional Kalman filter and smoother can then be implemented without additional approximations provided that we include all the spatial measurement and test points in the state vector [11]. See Example 4 for a demonstration of this.

E. Non-Causal Stochastic Partial Differential Equations

An important thing to realize is that even if the spatio-temporal covariance function was originally constructed as a solution to some kind of stochastic partial differential equation,

Example 4 (2D squared exponential covariance function)**.** *The squared exponential covariance function*

$$k(\mathbf{x}, t; \mathbf{x}', t') = \sigma^2 \exp(-\alpha\|\mathbf{x} - \mathbf{x}'\|^2 - \alpha|t - t'|^2),$$

where $\alpha = 1/(2l^2)$, is separable, which means that its spectral density

$$S(\boldsymbol{\omega}_x, \omega_t) = \left(\frac{\pi}{\alpha}\right)^{d/2} \exp\left(-\frac{\|\boldsymbol{\omega}_x\|^2}{4\alpha}\right) \exp\left(-\frac{\omega_t^2}{4\alpha}\right)$$

is also separable. We use the truncated series approximation to the temporal part from Example 2, which leaves us with an approximation $S(\boldsymbol{\omega}_x, \omega_t) \approx |G(i\,\omega_t)|^2 S_w(\boldsymbol{\omega}_x)$. If we define $\mathbf{f}(\mathbf{x}, t) = (f(\mathbf{x}, t), \partial f(\mathbf{x}, t)/\partial t, \ldots, \partial^{n-1} f(\mathbf{x}, t)/\partial t^{n-1})$, collecting the terms from the transfer function gives the solution

$$\frac{\partial \mathbf{f}(\mathbf{x}, t)}{\partial t} = \mathbf{A}\,\mathbf{f}(\mathbf{x}, t) + \mathbf{L}\,w(\mathbf{x}, t),$$

where $w(\mathbf{x}, t)$ is a time-white process with a spatial spectral density $S_w(\boldsymbol{\omega}_x)$ and the matrix $\mathbf{A}$ is the same as in Example 2. Because $\mathbf{A}$ does not contain any operators, the corresponding infinite-dimensional Kalman filter and smoother can be implemented without any additional approximations.

we might still need to do the above factorization. For example, consider the following stochastic partial differential equation (SPDE) due to Whittle [22]:

$$\frac{\partial^2 f(x,t)}{\partial x^2} + \frac{\partial^2 f(x,t)}{\partial t^2} - \lambda^2 f(x,t) = w(x,t), \tag{37}$$

where $w(x,t)$ is a space–time white Gaussian random field. Fourier transforming the system and computing the spectral density gives the stationary covariance function

$$C(x,t) = \frac{\sqrt{x^2 + t^2}}{2\lambda} K_1(\lambda\sqrt{x^2 + t^2}), \tag{38}$$

which can be seen as a special case of the covariance function in Example 3. But if we converted the SPDE in Equation (37) into a state space model, we would get a different state space model than in Equation (36) and the model would not even contain pseudo-differential operators at all. Now the catch is that if we did that, the resulting model would not be a stable system. This is because the Equation (37) corresponds to selection of the roots of the spectral density in such way that all of them are not in the upper half of the complex plane. Thus the process is not Markovian.

IV. Infinite-Dimensional Bayesian Filtering, Smoothing and Parameter Estimation

A. Infinite-Dimensional Kalman Filtering and Smoothing of Spatio-Temporal Gaussian Processes

Using the procedure outlined in the previous section we can convert given stationary spatio-temporal covariance functions into equivalent infinite-dimensional state space models. The spatio-temporal Gaussian process regression solution can be then computed with the infinite-dimensional Kalman filter and smoother [18], [11].

However, in practice, we cannot implement the infinite-dimensional Kalman filters and smoothers exactly, but the pseudo-differential operator equations appearing in the equations need to be solved numerically. Fortunately, we can use the wide arsenal of numerical methods developed for deterministic pseudo-differential and partial differential equation models for that purpose. Because stationary covariance function models always lead to equations, which can be expressed in terms of the Laplace operator, a particularly useful method is to approximate the solution using the eigenbasis of the Laplace operator [11], [23].

B. Non-Linear and Non-Gaussian Extension

It is also possible to extend the present methodology into non-Gaussian measurement models (e.g., classification problems). In principle, the only difference is that we just need to replace the infinite-dimensional Kalman filter update with the corresponding infinite-dimensional extension of a non-linear Kalman filter (cf. [24]). The resulting approximations are similar to what has previously been used in context of non-linear inverse problems [3], [14]. To some extent it would also be possible to construct non-Gaussian prior models by allowing non-linearities in the infinite-dimensional state space model. There indeed exists a mathematical theory for this kind of models [25], but the numerical treatment is quite challenging.

C. Parameter Estimation

Because after the conversion procedure, the model is a state space model, we can estimate the parameters in the model using the methods developed for estimation of parameters in finite-dimensional state space models. These methods also inherit the pleasant linear time complexity whereas Gaussian process regression based parameter estimation methods (e.g., [2]) also have the cubic complexity problem. A good review of the available methods can be found, for example, in the book [26]. Parameter estimation in SDE based models was also recently studied in [27] and methods also suitable for infinite-dimensional systems were recently discussed in [28].

V. Application Examples

A. Spatio-Temporal Modeling of Precipitation

As the first application example we consider spatio-temporal interpolation of precipitation levels based on monthly data collected during the years 1895–1997 at over 400 stations around Colorado, US. The same data was also used by [11] and it contains total number of over 300 000 observations. We used a Gaussian process model with the non-separable spatio-temporal Matérn covariance function in Example 3 with smoothness $\nu = 3/2$, and a 10-year subset of the data. In this case, the direct GP solution would require inversion of a $55\,410 \times 55\,410$-matrix, which renders inference with the model practically unusable. The filters and smoothers were approximated by using truncated eigenfunction expansion of the Laplace operator with 384 eigenfunctions, and the parameters were optimized with respect to marginal likelihood.

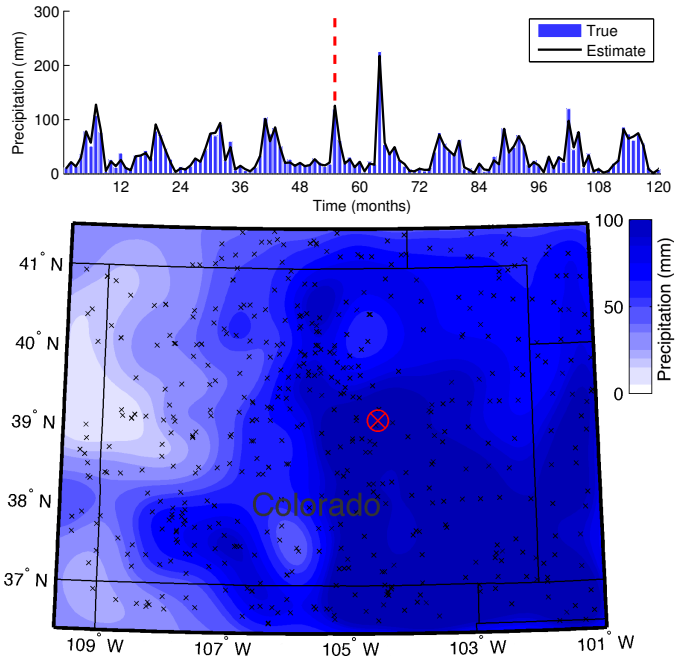


Fig. 4. Spatial map of precipitation levels for one month, and the temporal series for one location. The time of the spatial map is marked to the time series with red line, and the location of the time series on the spatial map is marked in red.

Figure 4 shows the resulting precipitation field for one month. The temporal time series for one location is presented for a test point that was left out from the estimation. As can been seen in the figure, the method provides good spatial interpolation of the data while still following the temporal variations quite well.

B. Oscillatory Structures in Brain Data

Instead of starting from a GP regression problem, we can also formulate the physical phenomena directly as an infinite-dimensional state space model and combine it with spatio-temporal covariance function models. This means that we can base the modeling on a wide range of first principles models. Following the presentation in [23], we form a SPDE model for spatio-temporal oscillators

$$\frac{\partial^2 f_j(\mathbf{x},t)}{\partial t^2} + \mathcal{A}_j \frac{\partial f_j(\mathbf{x},t)}{\partial t} + \mathcal{B}_j f_j(\mathbf{x},t) = \xi_j(\mathbf{x},t), \quad (39)$$

where the phenomena is modeled as a superposition of several latent components j. This model features three types of spatial dependency: operators $\mathcal{A}_j$ and $\mathcal{B}_j$ define spatial coupling and temporal oscillations, and spatio-temporal structure in the process noise term $\xi_j(\mathbf{x},t)$ can be included through a suitably chosen covariance function. In [23] the operators were chosen such that

$$\begin{aligned} \mathcal{A}_j &= \gamma_j \mathcal{I} - \chi_j \nabla^2 \\ \mathcal{B}_j &= \frac{1}{2}(\gamma_j - \chi_j \nabla^2)^2 + \omega_j^2, \end{aligned} \quad (40)$$

where $\gamma_j, \chi_j \geq 0$ are some non-negative constants affecting the damping. Here $\mathcal{I}$ is the identity operator and ω_j stands for the angular velocity (frequency). As a real-world example

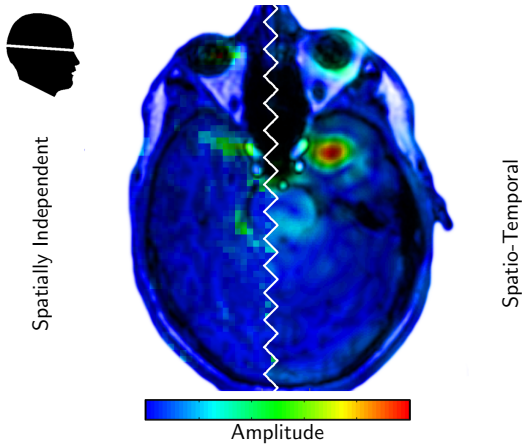


Fig. 5. Spatial amplitude map for heart beat induced noise in the brain estimated both using a spatially independent (left) and a spatio-temporal model (right). The slice orientation is shown on the left.

we demonstrate the estimation of the amplitude of heart beat induced oscillations in functional magnetic resonance imaging (fMRI) brain data from a healthy volunteer. The matrix size was 64×64, repetition time $0.1\,\mathrm{s}$ and total length $30\,\mathrm{s}$.

We use the known heart beat rate in forming the oscillator models. The estimation results in a spatio-temporal field that models the cardiac-induced physiological noise in the brain. We use this result to estimate a spatial map of the mean amplitude for the cardiac influence. Figure 5 shows the amplitude map of both the spatially independent estimation outcome and the spatio-temporal oscillator model. The hot spots in the figure correlates with the large blood vessels. In fMRI, the spatio-temporal model can mitigate the problems related to slow sampling, as the spatial information can compensate for missing temporal data.

REFERENCES

[1] A. E. Gelfand, P. J. Diggle, Montserrat, and P. Guttorp, *Handbook of Spatial Statistics*. CRC Press, 2010.
[2] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.
[3] A. Tarantola, *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM, 2004.
[4] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*. Springer, 2005.
[5] M. S. Grewal and A. P. Andrews, *Kalman Filtering, Theory and Practice Using MATLAB*. Wiley Interscience, 2001.
[6] Z. Chen and S. Haykin, "On different facets of regularization theory," *Neural Computation*, vol. 14, no. 12, pp. 2791–2846, 2002.
[7] B. Øksendal, *Stochastic Differential Equations: An Introduction with Applications*, 6th ed. Springer, 2003.
[8] M. Alvarez and N. D. Lawrence, "Latent force models," in *JMLR Workshop and Conference Proceedings Volume 5 (AISTATS 2009)*, 2009, pp. 9–16.
[9] J. Hartikainen, M. Seppänen, and S. 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 (ICML 2012)*, 2012.
[10] J. Hartikainen and S. Särkkä, "Kalman filtering and smoothing solutions to temporal Gaussian process regression models," in *Proceedings of IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2010.
[11] S. Särkkä and J. Hartikainen, "Infinite-dimensional Kalman filtering approach to spatio-temporal Gaussian process regression," in *JMLR Workshop and Conference Proceedings Volume 22 (AISTATS 2012)*, 2012, pp. 993–1001.
[12] F. Lindgren, H. Rue, and J. Lindström, "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 4, pp. 423–498, 2011.
[13] N. Cressie and C. K. Wikle, "Space-time Kalman filter," in *Encyclopedia of Environmetrics*, A. H. El-Shaarawi and W. W. Piegorsch, Eds. John Wiley & Sons, Ltd, Chichester, 2002, vol. 4, pp. 2045–2049.
[14] P. Hiltunen, S. Särkkä, I. Nissilä, A. Lajunen, and J. Lampinen, "State space regularization in the nonstationary inverse problem for diffuse optical tomography," *Inverse Problems*, vol. 27, p. 025009, 2011.
[15] L. Csató and M. Opper, "Sparse on-line Gaussian processes," *Neural Computation*, vol. 14, no. 3, pp. 641–668, 2002.
[16] W. Liu, I. Park, Y. Wang, and J. Príncipe, "Extended kernel recursive least squares algorithm," *IEEE Transactions on Signal Processing*, vol. 57, no. 10, pp. 3801–3814, 2009.
[17] S. Van Vaerenbergh, M. Lázaro-Gredilla, and I. Santamaría, "Kernel recursive least-squares tracker for time-varying regression," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 8, pp. 1313–1326, 2012.
[18] R. Curtain, "A survey of infinite-dimensional filtering," *SIAM Review*, vol. 17, no. 3, pp. 395–411, 1975.
[19] S. Särkkä, "Linear operators and stochastic partial differential equations in Gaussian process regression," in *Lecture Notes in Computer Science Volume 6792 (ICANN 2011)*, 2011, pp. 151–158.
[20] T. Glad and L. Ljung, *Control Theory: Multivariable and Nonlinear Methods*. Taylor & Francis, 2000.
[21] M. A. Shubin, *Pseudodifferential operators and spectral theory*. Springer-Verlag, 1987.
[22] P. Whittle, "On stationary processes in the plane," *Biometrica*, vol. 41, no. 3/4, pp. 434–449, 1954.
[23] A. Solin, "Hilbert space methods in infinite-dimensional Kalman filtering," Master's thesis, Aalto University, 2012.
[24] S. Särkkä and J. Sarmavuori, "Gaussian filtering and smoothing for continuous-discrete dynamic systems," *Signal Processing*, vol. 93, no. 2, pp. 500–510, 2013.
[25] G. Da Prato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions*. Cambridge University Press, 1992.
[26] O. Cappé, E. Eric Moulines, and T. Rydén, *Inference in hidden Markov models*. Springer, 2005.
[27] I. S. Mbalawata, S. Särkkä, and H. Haario, "Parameter estimation in stochastic differential equations with Markov chain Monte Carlo and non-linear Kalman filtering," *Computational Statistics*, 2013, (in press).
[28] H. Singer, "Continuous-discrete state-space modeling of panel data with nonlinear filter algorithms," *AStA Advances in Statistical Analysis*, vol. 95, no. 4, pp. 375–413, 2011.