Title
Dependent Matérn Processes for Multivariate Time Series

Permalink
https://escholarship.org/uc/item/3h87d0kj

Authors
Vandenberg-Rodes, Alexander
Shahbaba, Babak

Publication Date
2015-02-11

Peer reviewed

Dependent Matérn Processes for Multivariate Time Series

Alexander Vandenberg-Rodes VANDENBE@UCI.EDU
Department of Statistics, University of California, Irvine

Babak Shahbaba BABAKS@UCI.EDU
Department of Statistics, University of California, Irvine

Abstract

For the challenging task of modeling multivariate time series, we propose a new class of models that use dependent Matérn processes to capture the underlying structure of data, explain their interdependencies, and predict their unknown values. Although similar models have been proposed in the econometric, statistics, and machine learning literature, our approach has several advantages that distinguish it from existing methods: 1) it is flexible to provide high prediction accuracy, yet its complexity is controlled to avoid overfitting; 2) its interpretability separates it from black-box methods; 3) finally, its computational efficiency makes it scalable for high-dimensional time series. In this paper, we use several simulated and real data sets to illustrate these advantages. We will also briefly discuss some extensions of our model.

1. Introduction

Developing powerful models that can capture the dynamics of multivariate time series data, in order to explain their dependencies and predict their unknown values, remains a difficult task in statistics and machine learning. A key challenge is to answer:

Question. *How can we describe correlations* **among** *multiple time series*

$$x_1(t), x_2(t), \ldots, x_p(t), \qquad (1)$$

in a way that is also useful for prediction?

In this paper, we tackle this issue by proposing a special case of multivariate Gaussian processes that we call Dependent Matérn Processes (DMP). Similar models have been previously proposed in the econometrics, statistics, and machine learning literature. Here, we follow the recent work of (Sarkka et al., 2013) in considering Gaussian processes from the viewpoint of stochastic differential equations, and attempt to elucidate the mathematical underpinnings of this approach. Despite the similarity to several existing methods, our focus is on constructing a more interpretable model that can explain dependencies among multiple time series, but without sacrificing flexibility or scalability.

This paper is organized as follows. We discuss univariate Gaussian process (GP) models in Section 2, and briefly review several methods to generate multivariate GP. In Section 3, we present our proposed method, Dependent Matérn Processes. This is a special case of multivariate GP with properties that make it a powerful alternative to existing methods. Section 4 shows how univariate GPs can be in fact presented either as infinite dimensional functions or as solutions to a specific class of stochastic differential equations. Following (Sarkka et al., 2013), we show how these two alternative representations are connected, and use this insight to develop the inferential framework of our DMP model in Section 5. In Section 6, we present several experiments to illustrate the advantages of our method. Finally, in Section 7, we discuss possible extensions of this approach.

2. Preliminaries

Throughout this paper we will stay within the framework of Gaussian processes (GP). In this section, we discuss univariate and multivariate GP. We represent scalar quantities with lower-case and use capital letters to represent vectors. Boldface capital letters represent matrices.

2.1. Univariate Gaussian Processes

A Gaussian process (GP) on the real line is a random real-valued function $x(t)$, with statistics completely determined by its mean function $\mathbb{E}x(s)$ and *kernel* $\kappa(s,t) = \mathrm{Cov}(x(s), x(t))$. More precisely, all finite-dimensional distributions $(x(t_1), \ldots, x(t_n))$ are multivariate Gaussian with mean $(\mathbb{E}x(t_1), \ldots, \mathbb{E}x(t_n))$, and with covariance matrix $(\kappa(t_k, t_\ell))_{k,\ell=1}^n$. Since the latter must be positive semidefinite for every finite collection of inputs $t_1, \ldots, t_n$, only certain kernels κ are *valid* – that is, define Gaussian processes. Thus when using Gaussian processes, a practitioner often chooses from among the few popular classes of kernels, such as the Squared Exponential (SE), Ornstein-Uhlenbeck (OU), Matérn, Polynomial, and linear combinations of these.

In general, the choice of kernel encodes our *qualitative* beliefs about the underlying signal. For instance, the OU kernel produces non-differentiable functions $x(t)$, while the SE kernel is infinitely differentiable. In this paper we will concentrate on the Matérn class of kernels, which have as hyper-parameters the smoothness ν, length-scale ℓ, and variance σ^2. In particular, for $n = 0, 1, \ldots$ and $\nu = \frac{1}{2}+n$ it is known that realizations $x(t)$ of a GP with Matérn kernel are n times continously differentiable, while $\kappa(s,t)$ decays at rate $e^{-|t-s|\sqrt{2\nu}/\ell}$ as $t - s$ becomes large (Stein, 1999; Rasmussen & Williams, 2006). It is then straightforward to add extra *observation noise* reflecting our uncertainty in the accuracy of our measurements.

2.2. Multivariate GPs

There often arise situations where we would like to jointly model several time series $x_1(t), x_2(t), \ldots, x_p(t)$, for the purpose of *inference*, in particular attempting to quantify the dependencies between the observed series, and/or to improve our *predictions* of one series using data from the others. In the context of Gaussian processes, we intend that for different processes i and j we have a non-zero covariance. In fact, a *multi-output* or *multivariate* Gaussian process can be defined just as in Section 2.1, but where the kernel function now depends on two pairs of inputs. For simplicity we will assume in what follows that the mean of each time series is the zero function. The kernel κ is now defined for $i, j = 1, \ldots, p$ and $s, t \in \mathbb{R}$ as

$$\kappa([i,s],[j,t]) = \mathbb{E}x_i(s)x_j(t). \tag{2}$$

The initial challenge within the Gaussian process context is to produce a valid and interpretable kernel.

2.2.1. Linear models

The usual technique for generating multivariate GP kernels is known as *co-kriging* from the geostatistical literature (Cressie, 1993). The simplest case is known as the *intrinsic co-regionalization model* (ICM), where one takes $\mathbf{C} = (c_{ij})$ to be a positive definite $p \times p$ matrix, $\kappa^{(1)}(s,t)$ to be a valid univariate kernel, and defines the multi-output kernel κ to be their product

$$\mathbb{E}x_i(s)x_j(t) = c_{ij}\kappa^{(1)}(s,t). \tag{3}$$

Notice that while the intrinsic co-regionalization model is easily interpretable – the single matrix $\mathbf{C}$ provides the covariances between the time series – all outputs share the same univariate kernel, which makes for a rather inflexible model.

The *linear model of coregionalization* (LCM) adds more flexibility by allowing linear combinations of ICM's, resulting in a kernel of the form

$$\mathbb{E}x_i(s)x_j(t) = \sum_{k=1}^{q} c_{ij}^{(k)}\kappa^{(k)}(s,t). \tag{4}$$

For each $k = 1, \ldots, q$, $\kappa^{(k)}(s,t)$ is assumed to be a valid kernel for a univariate GP, and $\mathbf{C}^{(k)} = (c_{ij}^{(k)})$ is assumed to be a positive definite matrix. It is not hard to see that (4) results in a valid kernel. However, we have now lost some the interpretability of the ICM. More problematically, the LCM still does not provide a notion of correlation between processes with differing length-scales. One proposed solution is the *process convolution* approach (Boyle & Frean, 2005; Alvarez & Lawrence, 2011), which allows for qualitatively very different processes to be correlated, though with some loss of interpretability.

2.2.2. Latent models

Another approach is to describe $(x_1(t), \ldots, x_p(t))$ as linear combinations of *latent* factors. We suppose $u_1(t), \ldots, u_q(t)$ are independent mean zero Gaussian processes, and let

$$x_i(t) = \sum_{k=1}^{q} a_{i,k}u_k(t), \qquad \text{for } i = 1, 2, \ldots p. \tag{5}$$

Let $\kappa_i(s,t) = \mathbb{E}u_i(s)u_i(t)$ be the kernel for the i'th latent process. Then the observed processes $\mathbf{x}(t) = (x_1(t), \ldots, x_p(t))$ are jointly mean-zero Gaussian with covariances

$$\mathbb{E}x_i(s)x_j(t) = \sum_{k=1}^{q} a_{i,k}a_{j,k}\kappa_k(s,t). \tag{6}$$

This is the *semi-parametric latent factor model* of (Teh et al., 2005), so-called because the linear combination of latent GP's is parameterized by the matrix of coefficients $A = (a_{i,k})$, while each Gaussian process is of course a non-parametric model. However, this latent model (5) is actually an example of the above linear model of coregionalization, where $\mathbf{C}^{(k)}$ is just the outer product of the vector $a_{\cdot,k}$ with itself.

See (Alvarez et al., 2011) for a nice survey of these and other variants of co-kriging used in the machine learning literature.

2.2.3. Other approaches

Instead of trying to create multivariate kernels in a general fashion, one can attempt multivariate generalizations of a given class of univarate kernels, often by using Bôchner's theorem (see Section 4.1 below). The recent work of (Gneiting et al., 2010; Apanasovich et al., 2012) is perhaps the most relevant to our model, as they show how to construct a family of valid kernels for multivariate Gaussian processes on $\mathbb{R}^d$ where the marginal processes each have Matérn kernel with different hyperparameters.

3. Dependent Matérn processes

From a modeling perspective we would like to describe correlations between processes that have different (unique) hyperparameters, whereas in most of the above models this is only roughly attained by taking linear combinations of processes.

3.1. Our approach

We will model multivariate time series $X(t) = (x_1(t), \ldots, x_p(t))$ such that each marginal process $x_i(t)$ is a stationary mean-zero Gaussian process with Matérn kernel

$$\kappa_{\nu,\ell_j,\sigma_j}(t) = \mathbb{E}x_j(0)x_j(t),$$

thus the processes are allowed different length scales and variance, while sharing a common smoothness. In what follows we will always assume $n = \nu - \frac{1}{2}$ to be an integer. As we will explain in Section 4.3, each $x_j(t)$ can actually be represented as a solution of the stochastic differential equation

$$\left(\frac{d}{dt} + \frac{\sqrt{2\nu}}{\ell_j}\right)^{n+1} x_j(t) = \sigma_j C_{\nu,\ell_j} \dot{w}_j(t), \qquad (7)$$

where $\dot{w}(t)$ is white noise and C_{ν,ℓ_j} is a constant.

3.1.1. A new multi-output GP

To introduce dependence among the Matérn processes $x_j(t)$ we *correlate the input noises* $\sigma_j \dot{w}_j(t)$ in (7). That is, we let $\mathbf{L}$ be a $p \times R$ matrix and set

$$(w_1(t), \ldots, w_p(t))^T = diag(\sigma_1^{-1}, \ldots, \sigma_p^{-1})\mathbf{L}V(t), \quad (8)$$

where $V(t)$ is a vector of R independent standard Brownian motions, which we can think of as latent noise processes. Note that $\mathbf{L}$ has absorbed the σ_j parameters, and $(c_{ij}) = \mathbf{C} = \mathbf{L}\mathbf{L}^T$ is the covariance matrix of the input noises.

The stationary solution of these coupled SDEs results in multi-output GP, which we will refer to as a *Dependent Matérn process*.

In Section 5 we will show how to compute the kernel (2) for this new process, resulting in (for $\nu = \frac{1}{2}$)

$$\mathbb{E}x_i(s)x_j(t) \propto c_{ij} r_{ij} e^{-(t-s)/\ell_j}, \qquad (9)$$

while for $\nu = \frac{3}{2}$ we obtain

$$\mathbb{E}x_i(s)x_j(t) \propto c_{ij} r_{ij}^3 \frac{2 + (t-s)\left(\frac{\sqrt{3}}{\ell_i} + \frac{\sqrt{3}}{\ell_j}\right)}{e^{\sqrt{3}(t-s)/\ell_j}}. \qquad (10)$$

In both cases we are assuming $s \leq t$, and the factor

$$r_{ij} = 2\sqrt{\ell_i \ell_j}/(\ell_i + \ell_j) \qquad (11)$$

is the ratio of the geometric and arithmetic means of the two length-scales.

Examining these two expressions for the kernel, one should note:

1. They are not symmetric in time, as interchanging s and t would replace ℓ_j with ℓ_i in the exponential. That is, the covariance kernel respects the forward flow of time, which we believe to be a desired characteristic. Note this feature is missing from all of the models discussed above.
2. For $\ell_i \approx \ell_j$ the r_{ij} factor is close to 1, but as ℓ_i and ℓ_j increasingly differ in scale r_{ij} goes to zero. Intuitively this means that two processes with different length scales cannot move tightly together.

3.1.2. Defining the correlation

Even if the various length scales are quite different, the matrix $\mathbf{C} = (c_{ij})$, which we can recover from observed data, can be normalized in the usual way to obtain a clear, though model-dependent, notion of *correlation* (ρ_{ij}) *between time series*:

$$\begin{pmatrix} \rho_{11} & \cdots & \rho_{p1} \\ \vdots & \ddots & \\ \rho_{p1} & \cdots & \rho_{pp} \end{pmatrix} = \begin{pmatrix} c_{11}^{-\frac{1}{2}} & & \\ & \ddots & \\ & & c_{pp}^{-\frac{1}{2}} \end{pmatrix} C \begin{pmatrix} c_{11}^{-\frac{1}{2}} & & \\ & \ddots & \\ & & c_{pp}^{-\frac{1}{2}} \end{pmatrix}. \qquad (12)$$

3.1.3. Latent force models

Our proposed model can be thought of as a particular case of *latent force models* (Alvarez et al., 2009), although our motivation and approach to inference are very different. With a latent force model one thinks of each output time series as following specific physical dynamics, such as a damped harmonic oscillator, but that is also under the influence of latent forces (modelled as GPs), which are shared across the outputs as we do in (8).

With our model we are more interested in providing a *interpretable* notion of correlation between the time series, and do not assume knowledge of any underlying physical dynamics for the outputs. We are instead interested in the qualitative features of the Matérn class, and following, e.g. (Hartikainen & Sarkka, 2010; Mbalawata et al., 2013; Sarkka et al., 2013) we construct the SDE dynamics that represent such processes.

3.2. Computational complexity

Gaussian processes in general suffer from the *big-N* problem, that is, computations involving a N samples from a Gaussian process typically are of cubic complexity in N, since one usually needs to invert the $N \times N$ covariance matrix $(\kappa(t_i, t_j))$. In the case of p processes sampled N times, the resulting computational cost is $O(N^3 p^3)$, which can be already prohibitive when there are only a few hundred samples.

In special cases such as equally-spaced observations there are faster techniques such as circulant embedding (Dietrich & Newsam, 1997), and for general Gaussian processes there has been a recent flurry of research into sparse approximations (Quiñonero-Candela & Rasmussen, 2005).

As we will see in Section 4.2.2, stochastic differential equations can be transformed into state space models, which have the nice feature that computing the likelihood of N observations, or using the Kalman filter and Rauch-Tung-Streibel smoother for prediction, only has complexity $O(p^3 N)$. This allows our model to easily handle data containing thousands of observations.

In order to transform our DMP model into a state space model we first need to set up the mathematical background connecting Gaussian processes and stochastic differential equations. However, the reader might prefer to jump to Section 5.1, where we show how to use the state space form for inference and prediction.

4. Two approaches to univariate Gaussian processes

4.1. Infinite dimensional regression

One way of viewing a Gaussian process is as a random function of the form

$$x(t) = \sum_{k=0}^{\infty} a_k \psi_k(t), \tag{13}$$

where $\{\psi_k(t)\}$ is a collection of deterministic square integrable (L^2) functions, that is, *features*, and we place an iid $N(0,1)$ prior on the coefficients a_k. As a linear combination of Gaussians is Gaussian, $x(t)$ is clearly a GP.

To compute the kernel of (13), we take *any* orthonormal basis $\{\phi_n(t)\}$ of L^2, let g be an integrable function, and define $\psi_k(t)$ to be the convolution $\int \phi_k(u) g(t-u) du$, which should be thought of as the L^2 inner product of ϕ_k and $g(t - \cdot)$. Since $\mathbb{E} a_n a_m = \delta_{n,m}$, the kernel $\mathbb{E} x(s) x(t)$ reduces to

$$\sum_{k=0}^{\infty} \psi_k(s)\psi_k(t) = \int g(s-u) g(t-u) du. \tag{14}$$

The last equality is just Parsival's identity, relating the inner product of $g(s - \cdot)$ and $g(t - \cdot)$ to the dot product of their coefficient vectors $\{\psi_k(s)\}$ and $\{\psi_k(t)\}$ in the orthonormal basis. The key point of (14) – similar to the *kernel trick* for support vector machines – is that this kernel is independent of the choice of orthonormal basis, and so the function g now defines the Gaussian process.

By using the Fourier transform we can characterize the class of valid kernels. Given a *stationary* kernel ($\kappa(s,t) = \kappa(0, t-s)$), its *spectral density* $S(\xi)$ is defined by:

$$\kappa(0,t) = \int e^{it\xi} S(\xi) d\xi. \tag{15}$$

Noting that the right hand side of (14) describes a stationary kernel, we use Parsival's identity again to see that

$$\int g(t-u) g(-u) = \int e^{it\xi} |\hat{g}(\xi)|^2 d\xi, \tag{16}$$

with $\hat{g}$ the Fourier transform of g. In particular, a function $\kappa(t)$ with non-negative Fourier transform (spectral density) is a valid kernel for a stationary GP; the precise equivalence, known as Bôchner's theorem (Stein, 1999), shows that all valid kernels arise in this fashion.

4.2. Stochastic differential equations

A particularly nice way to construct Gaussian processes on the real line is via solutions of stochastic differential equations (SDE's). Although constructing Gaussian processes through SDE's goes back to the seminal article of (Doob, 1944), and has been used extensively in econometrics (Bergstrom, 1990), it has only recently seen development in the machine learning literature (Hartikainen & Sarkka, 2010; Hartikainen et al., 2012; Mbalawata et al., 2013; Sarkka et al., 2013; Reece et al., 2014).

The archtypical SDE is the Ornstein-Uhlenbeck process

$$\frac{dx}{dt}(t) = \alpha x(t) + \dot{w}(t), \tag{17}$$

where $\dot{w}(t)$ is Gaussian white noise. One can make mathematical sense of this via its integrated form

$$x(t) - x(s) = \int_s^t \alpha x(u) du + w(t) - w(s), \tag{18}$$

with $w(t)$ as the Weiner process (Brownian motion). Given an initial value $x(s)$, it has the solution

$$x(t) = e^{(t-s)\alpha} x(s) + \int_s^t e^{(t-u)\alpha} dw(u), \quad t \geq s. \tag{19}$$

Although it is sometimes thought that making sense of the integral in (19) requires the full weight of Itô calculus, for deterministic (and differentiable) integrands we can use the integration by parts formula: $\int_s^t f(u)dw(u) = w(t)f(t) - w(s)f(s) - \int_s^t f'(u)w(u)du$.[1]

4.2.1. General case

Higher order SDE's of the form

$$\frac{d^n}{dt^n}x(t) + a_{n-1}\frac{d^{n-1}}{dt^{n-1}}x(t) + \cdots + a_0 x(t) = \sigma\dot{w}(t), \quad (20)$$

such as the one defining the Matérn process (7), are similarly interpreted. Letting $F(t)$ be the vector of derivatives $(x(t), x'(t), \ldots, x^{(n-1)}(t))^T$, we can rewrite (20) as

$$\frac{dF}{dt}(t) = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ -a_0 & -a_1 & \cdots & -a_{n-1} \end{pmatrix} F(t) + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \sigma \end{pmatrix} \dot{w}(t). \quad (21)$$

With $\mathbf{Q}$ as the $n \times n$ matrix above and $E = (0, \ldots, 0, \sigma)^T$, the solution to (20) is completely analogous to (19):

$$F(t) = e^{(t-s)\mathbf{Q}}F(s) + \int_s^t e^{(t-u)\mathbf{Q}}E dw(u). \quad (22)$$

4.2.2. Stationarity

We now require that the eigenvalues of $\mathbf{Q}$, that is, the zeros of its characteristic polynomial

$$x^n + a_{n-1}x^{n-1} + \cdots + a_0, \quad (23)$$

all have negative real part. In this case, taking the limit of (22) as $t \to \infty$ results in a zero mean Gaussian random vector with a finite covariance matrix we denote by Σ_∞. If we then choose some initial point $F(0) \sim \mathcal{N}(0, \Sigma_\infty)$, the resulting process $(F(t);\ t \geq 0)$ is a stationary n-dimensional Gaussian Markov process, with covariance kernel

$$\mathbb{E}F(s)F(t)^T = e^{(t-s)\mathbf{Q}}\Sigma_\infty. \quad (24)$$

The integral in (22) is also Gaussian with covariance

$$\Sigma_\infty - e^{(t-s)\mathbf{Q}}\Sigma_\infty e^{(t-s)\mathbf{Q}^T}. \quad (25)$$

Usually we only observe the positions $x(t)$ at a finite collection of times $t_1, \ldots, t_N$. Assuming corruption by observation noise ϵ_k, the resulting observations of the SDE (20) can be written in the following state space form:

$$F(t_k) = e^{(t_k - t_{k-1})\mathbf{Q}}F(t_{k-1}) + \eta_k, \quad (26)$$

$$y(t_k) = HF(t_k) + \epsilon_k, \quad (27)$$

where $\{\eta_k\}$ are independent Gaussian with covariance (25), and $H = (1, 0, \ldots, 0)$ is the observation matrix.

4.3. Connecting SDEs to GPs

Unfortunately not all Gaussian processes on $\mathbb{R}$ exactly correspond to an SDE. The precise relationship is due to (Doob, 1944): A stationary Gaussian process on $\mathbb{R}$ can be represented as the stationary solution of (20) when its spectral density has the form

$$S(\xi) = \frac{\sigma^2}{|(i\xi)^n + a_{n-1}(i\xi)^{n-1} + \cdots + a_1(i\xi) + a_0|^2}. \quad (28)$$

The fundamental example is the Matérn class of Gaussian processes, which have a kernel with spectral density

$$S(\xi) = \sigma^2 C_{\nu,\ell}^2 \left(\xi^2 + \frac{2\nu}{\ell^2}\right)^{-(\nu+\frac{1}{2})}, \quad (29)$$

where ν, ℓ, and σ are the smoothness, lengthscale, and variance parameters, respectively, and $C_{\nu,\ell}$ is a constant with respect to ξ and σ. When $\nu = n + \frac{1}{2}$ we can factor $S(\xi) = \hat{g}(\xi)\overline{\hat{g}(\xi)}$, where[2]

$$\hat{g}(\xi) = \sigma C_{\nu,\ell}\left(i\xi + \frac{\sqrt{2\nu}}{\ell}\right)^{-n-1}. \quad (30)$$

Hence such Matérn class GP's can be realized as solutions of the SDE (7) used in our multivariate GP.

In order to put (7) into the state space form (26), we expand out its left hand side using the binomial theorem to obtain the $n+1 \times n+1$ matrix $\mathbf{Q}$ in (21). With $n = 0$ or 1 (corresponding to $\nu = 1/2$ or $3/2$), we have

$$\mathbf{Q}_j = (1/\ell_j), \text{ or } \mathbf{Q}_j = \begin{pmatrix} 0 & 1 \\ -\frac{3}{\ell_j^2} & -\frac{2\sqrt{3}}{\ell_j} \end{pmatrix}. \quad (31)$$

Furthermore, each matrix exponential can be computed analytically (Jones, 1981). With $n = 1$, for example, we have

$$e^{t\mathbf{Q}_j} = e^{-t\sqrt{3}/\ell_j}\begin{pmatrix} 1 + \frac{t\sqrt{3}}{\ell_j} & t \\ -\frac{3t}{\ell_j^2} & 1 - \frac{t\sqrt{3}}{\ell_j} \end{pmatrix}. \quad (32)$$

Although we will not make use of it here, one should note that by approximating $\hat{g}$ (16) with rational functions, one can approximately represent other Gaussian processes in terms of SDEs (Sarkka et al., 2013; Solin & Särkkä, 2014)

5. Inference in the dependent Matérn model

To obtain the *joint* state space representation of the p coupled SDE's (7), we stack the p derivative vectors

[1] In this interpretation only an interchange of integrals is required to show that (19) solves (18).

[2] There are multiple choices for $\hat{g}(\xi)$, however, only this one ensures that the zeros of $1/\hat{g}(ix)$ (that is, the polynomial (23)) all have negative real part, and thus corresponds to a stationary SDE as discussed in Section 4.2.2.

$F_1, \dots, F_p$ together to create the length $p(n+1)$ vector

$$\vec{F}(t) = (x_1(t), \dots, x_1^{(n)}(t), \dots, x_p(t), \dots, x_p^{(n)}(t))^T, \tag{33}$$

containing the p processes and their first n derivatives.

Recalling (26) and (27), write $\mathbf{E} = E \otimes \mathbf{I}_p$, $\mathbf{H} = H \otimes \mathbf{I}_p$, where $\mathbf{I}_p$ is the $p \times p$ identity matrix and $\mathbf{A} \otimes \mathbf{B}$ is the Kronecker product of the matrices $\mathbf{A}$ and $\mathbf{B}$. In particular, $\mathbf{H}\vec{F}(t) = (x_1(t), \dots, x_p(t))$. Then with $\vec{\mathbf{Q}}$ as the block diagonal matrix with blocks $\mathbf{Q}_1, \dots, \mathbf{Q}_p$ as in (31), the equivalent state space formulation of the coupled SDE's (7) can be written as

$$\vec{F}(t_k) = e^{(t_k - t_{k-1})\vec{\mathbf{Q}}}\vec{F}(t_{k-1}) + \vec{\eta}_k \tag{34}$$

$$Y(t_k) = \mathbf{H}\vec{F}(t_k) + \vec{\epsilon}_k. \tag{35}$$

Note that the matrix exponential is just a block diagonal matrix with blocks $e^{-\Delta t_k \mathbf{Q}_j}$. The observation noise $\vec{\epsilon}_k$ is assumed to be iid mean-zero Gaussian with diagonal covariance $diag(\tau_1^2, \dots, \tau_p^2)$. And finally the process noise $\vec{\eta}_k$ is given by (25). We omit the calculation of the needed stationary covariance Σ_∞ of $\vec{F}(t)$, which is a block matrix with the i, j-block an $n+1 \times n+1$ matrix

$$B_{ij} = c_{ij} r_{ij}, \quad \text{if } n = 0, \tag{36}$$

where r_{ij} was defined in (11), and when $n = 1$:

$$B_{ij} = c_{ij} r_{ij}^3 \begin{pmatrix} 2 & \frac{\sqrt{3}}{\ell_i} - \frac{\sqrt{3}}{\ell_j} \\ \frac{\sqrt{3}}{\ell_j} - \frac{\sqrt{3}}{\ell_i} & \frac{6}{\ell_i \ell_j} \end{pmatrix}. \tag{37}$$

Finally, by substituting (36) and (37) into (24), we can obtain the covariances (9) and (10).

5.1. Applying the Kalman filter and smoother

Given a state space model

$$z_k = A_k z_{k-1} + \eta_k,$$
$$y_k = H_k z_k + \epsilon_k,$$

and observed data $y_1, \dots, y_N$, with η_k and ϵ_k as independent Gaussian noise, the Kalman filter (see Murphy (2012) for example) recursively calculates the conditional means and covariances

$$m_k^- = \mathbb{E}(z_k | \, y_1, \dots, y_{k-1}, \Theta) \tag{38}$$

$$P_k^- = \mathbb{E}\left((z_k - m_k^-)(z_k - m_k^-)^T | y_1, \dots, y_{k-1}, \Theta\right). \tag{39}$$

We use Θ to denote the collected parameters for A_k, η_k, and ϵ_k. Setting $S_k = H_k P_k^- H_k^T + J_k$, where J_k is the covariance matrix of the observation noise ϵ_k, the log likelihood $\log \mathbb{P}(\Theta | y_1, \dots, y_N)$ is, up to a constant,

$$\begin{aligned} &\log \mathbb{P}(\Theta) + \sum_{k=1}^N \log \mathbb{P}(y_k | y_1, \dots, y_{k-1}, \Theta) \\ &= \log \mathbb{P}(\Theta) + \sum_{k=1}^N \log \mathcal{N}(y_k; H_k m_k^-, S_k). \end{aligned} \tag{40}$$

For prediction we can use the Rauch-Tung-Streibel smoother to obtain the means and covariances,

$$m_{k;N} = \mathbb{E}(z_k | y_1, \dots, y_N, \Theta), \tag{41}$$

$$P_{k;N} = \mathbb{E}\left((z_k - m_{k;N})(z_k - m_{k;N})^T | y_1, \dots, y_N, \Theta\right), \tag{42}$$

conditional on the training data and the inferred parameters Θ. Note that the state space framework easily handles the missing (test) data by modifying the observation matrix H_k.

5.2. Implementation

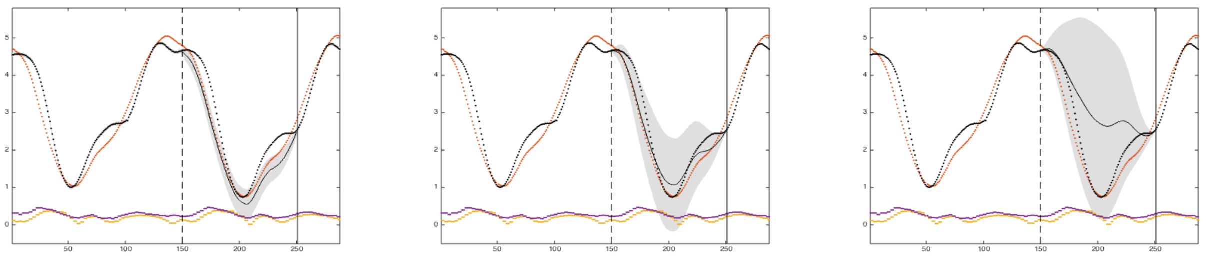
In the case of our state space model (34) and (35), we assume that the smoothness ν, and the number of latent noise processes R is chosen ahead of time. Hence our collected parameters Θ are: $\ell_1, \dots, \ell_p$ (length-scale parameters), $\mathbf{L}$ (a $p \times R$ matrix parameterizing the covariance across the observed processes), and $\tau_1^2, \dots, \tau_p^2$ (variances of the observation noise). Our inference involves two stages:

1. Taking the state space form (26), (27), of each *univariate* Matérn process $x_j(t)$, we estimate the individual length-scales ℓ_j one-by-one by minimizing (40). In practice, we found that Matlab's fminunc() works well.

2. Now using the state space form (34), (35) for the multi-output process, we sample from the posterior distribution of the remaining parameters $\mathbf{L}$ and $\vec{\tau}^2$, using the Metropolis-Hastings algorithm.

6. Experiments

In this section, we use simulated and real data to evaluate our method. We compare our method to some existing algorithms in terms of prediction accuracy. Additionally, we show how our method describes correlations among multiple time series.



(a) $\nu = \frac{1}{2}$, SMSE: 0.059 (b) $\nu = \frac{3}{2}$, SMSE: 0.022 (c) $\nu = \frac{5}{2}$, SMSE: 0.224

Figure 2. Wave and Tide data– The values of Sotonmet tide heights between the two vertical lines are assumed to be unknown. The black dots represent the true values of the Sotonmet tide heights, and the black lines show the predicted mean, with ±2 standard deviations shaded, using three choices of smoothness parameters ν in our model. The standardized mean squared errors (SMSE) are provided for each option.

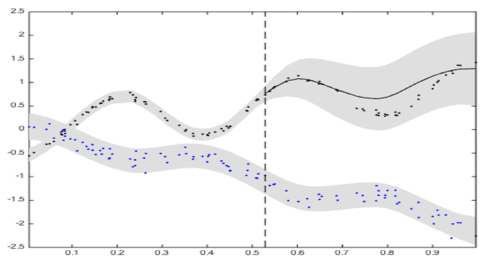


Figure 1. Simulated time series: $x_1(t) = 0.2\cos(5\pi t) - 2t + 0.1\epsilon$ and $x_2(t) = t - 0.5\cos(5\pi t) + 0.04\eta$ for $t \in [0, 1]$. The observed samples are shown with blue and black dots, respectively. The black line illustrates the Kalman-filter predicted means for the withheld samples.

6.1. Synthetic data

For our first experiment, we took a random selection of 100 times $t \in [0, 1]$, and simulated two time series

$$x_1(t) = 0.2\cos(5\pi t) - 2t + 0.1\epsilon,$$
$$x_2(t) = t - 0.5\cos(5\pi t) + 0.04\eta,$$

where ϵ and η are iid $\mathcal{N}(0, 1)$ noises. These are shown as the blue and black dots, respectively, in Figure 1. We removed the last 41 observations of the second time series (black dots), illustrated by the vertical dashed line, and treated them as the test set. The black line shows the Kalman filter predicted means for the withheld data, using the last sampled parameters based on our model, and the gray area shows the given $\pm 2\sigma$ deviations about the predicted mean for both series. On a 2011 Macbook with a 2.3Ghz i5 processor and 8GBs of RAM it took 65 seconds to draw 50,000 posterior samples of the correlation and noise parameters, while estimating the length scales and predicting the missing values is near-instantaneous. A highly optimized Kalman filter routine might lower the sampling time by an order of magnitude.

6.2. Wave and Tide data

For our second experiment, we tested our model on wave and tide data from the weather stations of Cambermet, Chimet, and Sotonmet, all on the southern coast of the U.K.[3]. The data consists of four time series: the tide heights of Chimet and Sotonmet, and the wave heights of Cambermet and Chimet. There are 288 observations, taken at 5 minute intervals, from the day of January 1, 2010. Observations 150 to 250 of the Sotonmet tide heights (black dots) were removed to make a test set.

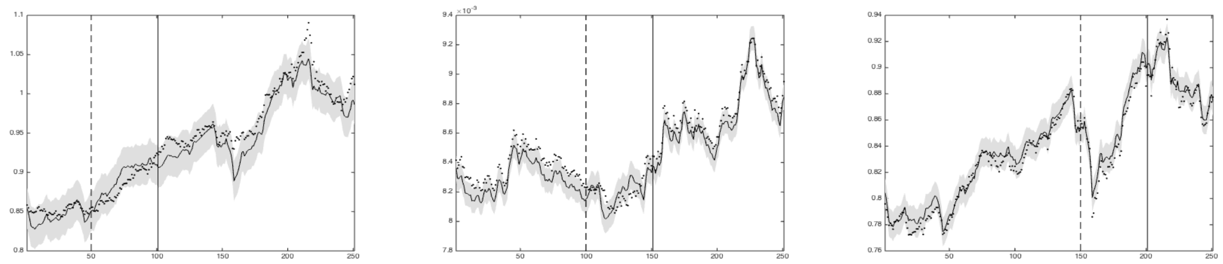
With this data we investigated how different choices of the smoothness parameter ν affected performance. All simulations used $R = 4$ independent noise sources. In figure 2 the black dots represent the true values of the Sotonmet tide heights, and the black line is the predicted mean, with ±2 standard deviations shaded.

With $\nu = 1/2$ the model overestimates the correlation between the two tide heights ($\rho \approx 0.9$), resulting in an overconfident estimate that tracks the other tide height (red dots) too closely. With $\nu = 5/2$ we have the opposite problem: despite the two tide heights staying together over the course of the day, there is not much correlation found between their third derivatives, resulting in a very weak prediction. The middle case of $\nu = 3/2$ strikes a nice balance, with estimated length-scales of $(102.5, 75.4, 24.0, 41.0)$, and inferred correlation matrix

$$\begin{pmatrix} 1.0000 & 0.6155 & 0.0191 & 0.0655 \\ 0.6155 & 1.0000 & 0.0547 & 0.0984 \\ 0.0191 & 0.0547 & 1.0000 & 0.3344 \\ 0.0655 & 0.0984 & 0.3344 & 1.0000 \end{pmatrix}$$

Note the moderate correlations (≈ 0.6) found between the tide heights, and the weak correlations (< 0.1) between the

[3] Data available from http://www.chimet.co.uk



(a) CAD

(b) JPY

(c) AUD

Figure 3. The US Dollar exchange rate with respect to Canadian Dollar (CAD), Japanese Yen (JPY), and Australian Dollar (AUD). The black dots show the observed data. The vertical lines show the intervals where the data are assumed to be unknown (i.e., test set). The solid lines show the predicted means using our model. The grey areas show the corresponding 95% intervals.

tide and wave heights.

6.3. Financial data

For our last example, we consider the inference of missing data in the multivariate financial dataset used in (Alvarez et al., 2010). It contains thirteen time-series for the US Dollar exchange rate with respect to the top 10 international currencies (Canadian Dollar, Euro, Japanese Yen, Great British Pound, Swiss Franc, Australian Dollar, Hong Kong Dollar, New Zealand Dollar, South Korean Won, Mexican Peso), and three precious metals (gold, silver, platinum), over all 251 working days of the 2007 calendar year. Following (Alvarez et al., 2010) we removed the mean and normalized each series to have unit variance, and removed a test set of 251 data points, covering days 50-100, 100-150, and 150-200, from the Canadian Dollar, Japanese Yen, and Australian Dollar series, respectively. The remaining 3051 data points were used as the training set. (There are already 59 missing data points from the precious metal series). These three time series are shown in Figure 3, along with the predicted means. As before, the vertical lines show the intervals where the test set data was withheld.

Because of the roughness of the paths we modeled the 13 time series as dependent Matérn($\nu = \frac{1}{2}$) processes, and restricted the parameter space by allowing for only $R = 4$ independent noise sources. The predicted means and the corresponding 95% intervals are shown as solid lines and shaded areas respectively in Figure 3. We then compared with the linear model of coregionalization (LMC) where the kernel (4) is a combination of two Matérn($\nu = \frac{1}{2}$) kernels, and $\mathbf{C}^{(1)}$ and $\mathbf{C}^{(2)}$ are both of rank 2. Our model's predictions had a standardized mean squared error (SMSE) of 0.087 (averaged across the three test outputs), while the LMC scored 0.49. Note that in this case our model is essentially the Stochastic Latent Force model in of (Alvarez et al., 2009), with all four latent processes as white noise. Nonetheless we end up with much better predictions (for their best model with one smooth and three white noise latent processes, (Alvarez et al., 2010) quote a SMSE of 0.2795, and 0.39 for their LMC implementation). We believe this shows the power of independently modelling the output processes and then their correlations.

7. Discussion

In this paper, we have proposed a new class of stochastic process models for multivariate time series. Using several examples, we illustrated our method's predictive power and interpretability. However, as discussed above, our method is also designed to be extendable to problems with more complex structures.

One possible extension to our model would be to allow kernels with (quasi-)periodic behavior, leading to better inference when modeling periodic phenomena such as the wave and tide data of Section 6.2. This is indeed possible within the state space approach, as exemplifed by the stochastic resonator model (Solin & Särkkä, 2013; 2014) and the linear basis model (Reece et al., 2014).

Referring again to the wave and tide data seen in Figure 2, one can see that the peaks and troughs are not perfectly aligned, either because of a delay in one of the sensor readings, or physical delay due to differing sensor locations. It should be possible to model this within the state space approach, allowing for more computationally efficient and interpretable versions of the Gaussian process sensor network model presented in (Osborne et al., 2012).

Acknowledgements

The first author would like to thank T. McClure for the many valuable discussions regarding Kalman filtering. This work is supported by NIH grant R01-AI107034.

References

Alvarez, Mauricio A and Lawrence, Neil D. Computationally efficient convolved multiple output gaussian processes. *The Journal of Machine Learning Research*, 12: 1459–1500, 2011.

Alvarez, Mauricio A, Luengo, David, and Lawrence, Neil D. Latent force models. In *International Conference on Artificial Intelligence and Statistics*, pp. 9–16, 2009.

Alvarez, Mauricio A., Luengo, David, Titsias, Michalis K., and Lawrence, Neil D. Efficient multioutput gaussian processes through variational inducing kernels. In Teh, Yee Whye and Titterington, D. Mike (eds.), *AISTATS*, volume 9 of *JMLR Proceedings*, pp. 25–32, 2010.

Alvarez, Mauricio A., Rosasco, Lorenzo, and Lawrence, Neil D. Kernels for Vector-Valued functions: a review. 2011. URL http://arxiv.org/abs/1106.6251.

Apanasovich, Tatiyana V, Genton, Marc G, and Sun, Ying. A valid matérn class of cross-covariance functions for multivariate random fields with any number of components. *Journal of the American Statistical Association*, 107(497):180–193, 2012.

Bergstrom, Albert Rex. *Continuous time econometric modelling*. Recent advances in econometrics. Oxford Univ. Press, 1990. ISBN 0198283407.

Boyle, Phillip and Frean, Marcus. Dependent gaussian processes. In *In Advances in Neural Information Processing Systems 17*, pp. 217–224. MIT Press, 2005.

Cressie, N. *Statistics for Spatial Data*. Wiley, New York, 1993.

Dietrich, C. and Newsam, G. Fast and exact simulation of stationary gaussian processes through circulant embedding of the covariance matrix. *SIAM Journal on Scientific Computing*, 18(4):1088–1107, 1997.

Doob, J L. The elementary gaussian processes. *The Annals of Mathematical Statistics*, 15(3):229–282, 1944.

Gneiting, Tilmann, Kleiber, William, and Schlather, Martin. Matérn Cross-Covariance functions for multivariate random fields. *Journal of the American Statistical Association*, 105(491):1167–1177, 2010.

Hartikainen, Jouni and Sarkka, Simo. Kalman filtering and smoothing solutions to temporal gaussian process regression models. pp. 379–384, 2010.

Hartikainen, Jouni, Seppanen, Mari, and Sarkka, Simo. State-space inference for non-linear latent force models with application to satellite orbit prediction. *arXiv preprint arXiv:1206.4670*, 2012.

Jones, R.H. Fitting a continous time autoregression to discrete data. *Applied Time Series Analysis II*, pp. 651–682, 1981.

Mbalawata, IsambiS., Srkk, Simo, and Haario, Heikki. Parameter estimation in stochastic differential equations with markov chain monte carlo and non-linear kalman filtering. *Computational Statistics*, 28(3):1195–1223, 2013. ISSN 0943-4062.

Murphy, Kevin P. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Osborne, Michael A, Roberts, Stephen J, Rogers, Alex, and Jennings, Nicholas R. Real-time information processing of environmental sensor network data using bayesian gaussian processes. *ACM Transactions on Sensor Networks (TOSN)*, 9(1):1, 2012.

Quiñonero-Candela, Joaquin and Rasmussen, Carl Edward. A unifying view of sparse approximate gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959, 2005.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning*. MIT Press, 2nd edition, 2006.

Reece, Steven, Ghosh, Siddhartha, Rogers, Alex, Roberts, Stephen, and Jennings, Nicholas R. Efficient state-space inference of periodic latent force models. *The Journal of Machine Learning Research*, 15(1):2337–2397, 2014.

Sarkka, Simo, Solin, Arno, and Hartikainen, Jouni. Spatiotemporal learning via Infinite-Dimensional bayesian filtering and smoothing: A look at gaussian process regression through kalman filtering. *IEEE Signal Processing Magazine*, 30(4):51–61, 2013.

Solin, Arno and Särkkä, Simo. Infinite-dimensional bayesian filtering for detection of quasiperiodic phenomena in spatiotemporal data. *Physical Review E*, 88(5): 052909, 2013.

Solin, Arno and Särkkä, Simo. Explicit link between periodic covariance functions and state space models. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33, pp. 904–912, 2014.

Stein, Michael L. *Statistical Interpolation of Spatial Data: Some Theory for Kriging*. Springer, 1999.

Teh, Y. W., Seeger, M., and Jordan, M. I. Semiparametric latent factor models. In *Proceedings of the International Workshop on Artificial Intelligence and Statistics*, volume 10, 2005.