SUPPLEMENTARY MATERIAL FOR "Infinite-Dimensional Kalman Filtering Approach to Spatio-Temporal Gaussian Process Regression"

1 Introduction

1.1 Wiener Process and White Noise

In the actual paper, we have denoted stochastic differential equations in Itô notation (cf. Karatzas and Shreve, 1991; Øksendal, 2003) such as

$$d\mathbf{f}(t) = \mathbf{A}\,\mathbf{f}(t)\,dt + \mathbf{L}\,d\mathbf{W}(t), \tag{1}$$

where $\mathbf{W}(t)$ is a Wiener process (or Brownian motion) with diffusion matrix $\mathbf{Q}_c$. The Wiener process is a Gaussian process with statistics:

$$\begin{aligned} \mathrm{E}[\mathbf{W}(t)] &= 0 \\ \mathrm{E}[\mathbf{W}(t)\,\mathbf{W}^T(s)] &= \mathbf{Q}_c\,\min(s,t)\,. \end{aligned} \tag{2}$$

In this supplementary material we will rewrite the equation (1) in differential equation form:

$$d\mathbf{f}(t)/dt = \mathbf{A}\,\mathbf{f}(t) + \mathbf{L}\,\mathbf{w}(t), \tag{3}$$

where the driving process $\mathbf{w}(t)$ is a Gaussian white noise with statistics

$$\begin{aligned} \mathrm{E}[\mathbf{w}(t)] &= 0 \\ \mathrm{E}[\mathbf{w}(t)\,\mathbf{w}^T(s)] &= \mathbf{Q}_c\,\delta(s-t), \end{aligned} \tag{4}$$

and can be considered as the formal derivative of Wiener process $\mathbf{w}(t) = d\mathbf{W}(t)/dt$. Here $\mathbf{Q}_c$ is called the spectral density of the white noise process. The space-time white noise can be defined in analogous manner.

The white noise notation is very convenient in practical computations, because in many cases the differential equations can be treated as if they were deterministic differential equations. For this reason this notation is often preferred in engineering literature (cf. Jazwinski, 1970; Grewal and Andrews, 2001). However, it is important to make sure that every operation is indeed valid in rigorous Itô calculus sense (Karatzas and Shreve, 1991; Øksendal, 2003), and treat the white noise notation only as a convenient notation for the actual Itô calculus in operation. To emphasis the actual meaning of the equations, we have chosen to use the Itô notation in the paper itself.

The background of the notation is that in rigorous sense, we cannot directly define differential equations driven by a white noise such as (3). Let's formally integrate the equation (3), which gives an integral equation of the form

$$\mathbf{f}(t) - \mathbf{f}(t_0) = \int_{t_0}^{t} \mathbf{A}\, \mathbf{f}(t)\, dt + \int_{t_0}^{t} \mathbf{L}\, \mathbf{w}(t)\, dt. \tag{5}$$

Now the last integral cannot be defined as Riemann integral, because the white noise process is formally non-continuous everywhere. However, it can be defined as so called Itô stochastic integral (see, e.g. Karatzas and Shreve, 1991; Øksendal, 2003) provided that we interpret the term $\mathbf{w}(t)\, dt$ as increment of Wiener process $\mathbf{W}(t)$. In Itô formalism the equation can be written in form

$$\mathbf{f}(t) - \mathbf{f}(t_0) = \int_{t_0}^{t} \mathbf{A}\, \mathbf{f}(t)\, dt + \int_{t_0}^{t} \mathbf{L}\, d\mathbf{W}, \tag{6}$$

where $d\mathbf{W}$ is the Wiener process increment. The second integral is now stochastic integral with respect to the stochastic "measure" $\mathbf{W}(t)$, the Wiener process. If we drop the integral signs and consider small values of $t - t_0$, the equation can be written in the more compact form (1), which is the most common notation for Itô stochastic differential equations in stochastics literature. The solution $\mathbf{f}(t)$ of an Itô stochastic differential equation is called an Itô process. Note that the equation can be formally written as

$$d\mathbf{f}(t)/dt = \mathbf{A}\, \mathbf{f}(t) + \mathbf{L}\, d\mathbf{W}/dt, \tag{7}$$

and comparing to Equation (3) reveals that the white noise process can be considered as the formal derivative of Wiener process $d\mathbf{W}/dt$. However, a slightly problematic thing is that the Wiener process is everywhere non-differentiable, and this causes appearance of the delta function in the covariance of white noise.

For the above reasons we also use the Itô notation for infinite-dimensional stochastic differential equations in the actual paper, because there the situation is analogous to the finite-dimensional case. In this supplement we use the white noise notation, because it is easier when doing the actual analytic calculations.

1.2 Multi-Dimensional Fourier Transform

The Fourier transform of function $f(\mathbf{x}) : \mathbb{R}^d \mapsto \mathbb{R}$ is here defined as

$$\mathcal{F}[f](\boldsymbol{\omega}) = \int_{\mathbb{R}^d} f(\mathbf{x}) \exp(-\mathrm{i}\, \boldsymbol{\omega}^T \mathbf{x})\, d\mathbf{x}. \tag{8}$$

The inverse transform is

$$\mathcal{F}^{-1}[F](\mathbf{x}) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} F(\boldsymbol{\omega}) \exp(\mathrm{i}\, \boldsymbol{\omega}^T \mathbf{x})\, d\boldsymbol{\omega}. \tag{9}$$

where $F(\boldsymbol{\omega}) = \mathcal{F}[f](\boldsymbol{\omega})$. Fourier transforms are rarely explicitly computed, but precomputed tables are often used instead (see, e.g. Råde and Westergren, 2004). One-dimensional Fourier transform pairs have been extensively tabulated in literature and because $\exp(\pm \mathrm{i}\, \boldsymbol{\omega}^T \mathbf{x}) = \prod_j \exp(\pm \mathrm{i}\, \omega_j\, x_j)$ multi-dimensional Fourier transforms can be computed as sequential application of single-dimensional transforms. The Fourier transform of a vector valued function can be computed by applying Fourier transform to each of the components of the vector separately.

The important properties, which make Fourier transform particularly useful for solving linear ordinary and partial differential equations are the following:

- Linearity: If $f(\mathbf{x})$ and $g(\mathbf{x})$ are arbitrary functions and $a, b \in \mathbb{R}$ are constants, then:
$$\mathcal{F}[a\, f + b\, g] = a\, \mathcal{F}[f] + b\, \mathcal{F}[g]. \tag{10}$$

- Derivative: If $f(\mathbf{x})$ is a k times differentiable function, defined on whole space $\mathbb{R}^d$ and vanishing at infinity, then the Fourier transform of the partial derivative $\partial^k f / \partial x_i^k$ is
$$\mathcal{F}[\partial^k f / \partial x_i^k] = (\mathrm{i}\, \omega_i)^k\, \mathcal{F}[f]. \tag{11}$$
That is, the Fourier transform maps derivatives to polynomials and thus transforms ordinary and partial differential equations into algebraic equations.

- Convolution: The convolution of functions $f(\mathbf{x})$ and $g(\mathbf{x})$ defined on whole space $\mathbb{R}^d$ as above can be defined as
$$(f * g)(\mathbf{x}) = \int_{R^d} f(\mathbf{x} - \mathbf{x}')\, g(\mathbf{x}')\, d\mathbf{x}'. \tag{12}$$
The Fourier transform of the convolution is then the product of Fourier transforms of f and g:
$$\mathcal{F}[f * g] = \mathcal{F}[f]\, \mathcal{F}[g]. \tag{13}$$

The Fourier transform is also useful in computing the covariance functions of stochastic ordinary and partial differential equations due to the following properties:

- Wiener-Khinchin: If $f(\mathbf{x})$ is a zero mean wide sense stationary random field with covariance function
$$C_f(\mathbf{u}) = \mathrm{E}[f(\mathbf{x})\, f(\mathbf{x} + \mathbf{u})], \tag{14}$$
then the spectral density $S_f(\boldsymbol{\omega})$ of the process $f(\mathbf{x})$ is the Fourier transform of $C_f(\mathbf{u})$:
$$S_f(\boldsymbol{\omega}) = \mathcal{F}[C_f]. \tag{15}$$

- If $h(\mathbf{x})$ is a function and $H(\mathrm{i}\boldsymbol{\omega})$ is Fourier transform (i.e., the transfer function), then the spectral density of the convolution process $g(\mathbf{x}) = h(\mathbf{x}) * f(\mathbf{x})$ is

$$S_g(\boldsymbol{\omega}) = H(\mathrm{i}\boldsymbol{\omega})\, S_f(\boldsymbol{\omega})\, H(-\mathrm{i}\boldsymbol{\omega}) = |H(\mathrm{i}\boldsymbol{\omega})|^2\, S_f(\boldsymbol{\omega}). \tag{16}$$

The Gaussian spatial white noise process can be defined as a random field $w(\mathbf{x})$ with the properties:

$$\begin{aligned} \mathrm{E}[w(\mathbf{x})] &= 0 \\ \mathrm{E}[w(\mathbf{x})\, w(\mathbf{x}+\mathbf{u})] &= q\, \delta(\mathbf{u}). \end{aligned} \tag{17}$$

The spectral density of the white noise process can be obtained as the Fourier transform of the covariance function $C_w(\mathbf{u}) = q\, \delta(\mathbf{u})$ and it is given as

$$S_w(\boldsymbol{\omega}) = q. \tag{18}$$

Due to this property the parameter q or its matrix equivalent in the definition of white noise is often called the spectral density of the white noise process.

In this document and in the paper write we stationary covariance function $C(\mathbf{x},\mathbf{x}') = C(\mathbf{x}-\mathbf{x}')$ simply as $C(\mathbf{x})$. In the case of spatio-temporal covariances, the stationary covariance functions are denoted as $C(\mathbf{x},t)$.

2 Details of Squared Exponential Covariance Function Example

The squared exponential (or exponential of square) class of covariance functions has the form

$$C(\mathbf{x}) = \exp\left(-\alpha\, \mathbf{x}^2\right), \tag{19}$$

where in the parameterization of Rasmussen and Williams (2006) we have $\alpha = 1/(2L^2)$. If we rename one of the input as t, and use separate scales for time and input, we get

$$\begin{aligned} C(\mathbf{x},t) &= \exp\left(-\alpha_x\, \mathbf{x}^2 - \alpha_t\, t^2\right) \\ &= \exp\left(-\alpha_x\, \mathbf{x}^2\right)\, \exp\left(-\alpha_t\, t^2\right) \end{aligned} \tag{20}$$

which can be seen to be separable in space and time. The corresponding spectral density is also separable

$$S(\boldsymbol{\omega}_x, \omega_t) = \left(\frac{\pi}{\alpha_x}\right)^{d/2} \exp\left(-\frac{\boldsymbol{\omega}_x^2}{4\alpha_x}\right) \left(\frac{\pi}{\alpha_t}\right)^{1/2} \exp\left(-\frac{\omega_t^2}{4\alpha_t}\right) \tag{21}$$

Following the procedure presented by Hartikainen and Särkkä (2010) we can now approximate the last term with a polynomial in ω_t^2:

$$\exp\left(-\frac{\omega_t^2}{4\alpha_t}\right) \approx \frac{1}{a_0 + a_1\,(i\omega_t)^2 + \cdots + a_N\,(i\omega)^{2N}}. \tag{22}$$

We can then form the spectral factorization, which will gives

$$\frac{1}{a_0 + a_1 (i\omega_t)^2 + \cdots + a_N (i\omega)^{2N}} = \underbrace{\left(\frac{1}{b_0 + b_1 (i\omega_t) + \cdots + b_N (i\omega_t)^N} \right)}_{H_t(i\omega_t)} \underbrace{\left(\frac{1}{b_0 + b_1 (-i\omega_t) + \cdots + b_N (-i\omega_t)^N} \right)}_{H_t(-i\omega_t)} \tag{23}$$

where $H_t(i\omega_t)$ has poles only in the upper half plane. Thus we get the approximation

$$S(\boldsymbol{\omega}_x, \omega_t) \approx \hat{S}(\boldsymbol{\omega}_x, \omega_t) = |H_t(i\omega_t)|^2 \, S_x(\boldsymbol{\omega}_x), \tag{24}$$

where

$$S_x(\boldsymbol{\omega}_x) = \left(\frac{\pi}{\alpha_x} \right)^{d/2} \left(\frac{\pi}{\alpha_t} \right)^{1/2} \exp\left(-\frac{\boldsymbol{\omega}_x^2}{4\alpha_x} \right). \tag{25}$$

Let $\boldsymbol{\omega}_x$ be fixed and consider the process $\tilde{f}$ satisfying the stochastic differential equation

$$b_0 \, \tilde{f}(\boldsymbol{\omega}_x, t) + b_1 \frac{\partial \tilde{f}(\boldsymbol{\omega}_x, t)}{\partial t} + \cdots + b_N \frac{\partial^N \tilde{f}(\boldsymbol{\omega}_x, t)}{\partial t^N} = \tilde{w}(\boldsymbol{\omega}_x, t), \tag{26}$$

where $t \mapsto \tilde{w}(\boldsymbol{\omega}_x, t)$ is a white noise process with spectral density $S_x(\boldsymbol{\omega}_x)$. The process now has the spectral density, which was defined in the Equation (24). Taking inverse Fourier transform with respect to the space then implies that the process satisfying the stochastic equation

$$b_0 \, f(\mathbf{x}, t) + b_1 \frac{\partial f(\mathbf{x}, t)}{\partial t} + \cdots + b_N \frac{\partial^N f(\mathbf{x}, t)}{\partial t^N} = w(\mathbf{x}, t), \tag{27}$$

where $w(\mathbf{x}, t)$ is a time-white process with spatial spectral density (25), and thus exponential covariance function, has the spectral density (24) and thus approximately the covariance function (20).

If we define $\mathbf{f} = (f, \partial f/\partial t, \ldots, \partial^{N-1} f/\partial t^{N-1})$, it is easy to see that the above equation can be written in form

$$\frac{\partial \mathbf{f}(\mathbf{x}, t)}{\partial t} = \mathbf{A} \, \mathbf{f}(\mathbf{x}, t) + \mathbf{L} \, w(\mathbf{x}, t) \tag{28}$$

where $\mathbf{A}$ and $\mathbf{L}$ are constant matrices.

3 Details of the Cressie & Huang Example

Consider the stationary covariance function introduced in Example 1 of Cressie and Huang (1999):

$$C(\mathbf{x}, t) = \frac{\sigma^2}{(a^2 t^2 + 1)^{d/2}} \exp\left(-\frac{b^2 ||\mathbf{x}||^2}{a^2 t^2 + 1} \right). \tag{29}$$

The spectral density is Gaussian in space and thus we get the spatial Fourier transform easily:

$$\begin{aligned}\mathcal{F}_x[C(\mathbf{x},t)] &= \frac{\sigma^2 \pi^{d/2}}{b^d} \exp\left(-\frac{a^2 t^2 + 1}{4b^2} ||\boldsymbol{\omega}_x||^2\right) \\ &= \frac{\sigma^2 \pi^{d/2}}{b^d} \exp\left(-\frac{||\boldsymbol{\omega}_x||^2}{4b^2}\right) \exp\left(-\frac{a^2 ||\boldsymbol{\omega}_x||^2}{4b^2} t^2\right).\end{aligned} \tag{30}$$

Taking Fourier transform with respect to t is again a Gaussian transform for the last term, which gives the spectral density

$$\begin{aligned}S(\boldsymbol{\omega}_x, \omega_t) &= \frac{\sigma^2 \pi^{d/2}}{b^d} \exp\left(-\frac{||\boldsymbol{\omega}_x||^2}{4b^2}\right) \left(\frac{2b\,\pi^{1/2}}{a\,||\boldsymbol{\omega}_x||}\right) \exp\left(-\frac{b^2}{a^2 ||\boldsymbol{\omega}_x||^2} \omega_t^2\right) \\ &= \frac{2\sigma^2 \pi^{(d+1)/2}}{a\,||\boldsymbol{\omega}_x||\,b^{d-1}} \exp\left(-\frac{||\boldsymbol{\omega}_x||^2}{4b^2}\right) \exp\left(-\frac{b^2}{a^2 ||\boldsymbol{\omega}_x||^2} \omega_t^2\right).\end{aligned} \tag{31}$$

Let's form the following Taylor series approximation to the inverse of the last term, write it in terms of $i\omega_t$ and factor out the highest order term:

$$\begin{aligned}&\exp\left(\frac{b^2}{a^2 ||\boldsymbol{\omega}_x||^2} \omega_t^2\right) \\ &\approx 1 + \left(\frac{b^2}{a^2 ||\boldsymbol{\omega}_x||^2}\right) \omega_t^2 + \frac{1}{2} \left(\frac{b^2}{a^2 ||\boldsymbol{\omega}_x||^2}\right)^2 \omega_t^4 \\ &= 1 - \left(\frac{b^2}{a^2 ||\boldsymbol{\omega}_x||^2}\right) (i\omega_t)^2 + \frac{1}{2} \left(\frac{b^2}{a^2 ||\boldsymbol{\omega}_x||^2}\right)^2 (i\omega_t)^4 \\ &= \frac{1}{2} \left(\frac{b^2}{a^2 ||\boldsymbol{\omega}_x||^2}\right)^2 \left(2 \left(\frac{a^2 ||\boldsymbol{\omega}_x||^2}{b^2}\right)^2 - 2 \left(\frac{a^2 ||\boldsymbol{\omega}_x||^2}{b^2}\right) (i\omega_t)^2 + (i\omega_t)^4\right)\end{aligned} \tag{32}$$

The roots of the polynomial on the right are given as

$$r = \pm 2^{1/4} \exp(\pm i\pi/8)\,||\boldsymbol{\omega}_x||\,(a/b), \tag{33}$$

and thus the stable roots are

$$r_s = -2^{1/4} \exp(\pm i\pi/8)\,||\boldsymbol{\omega}_x||\,(a/b). \tag{34}$$

By expanding the corresponding polynomial, we get the following:

$$(i\omega_t)^2 + 2^{5/4} \cos(\pi/8)\,||\boldsymbol{\omega}_x||\,(a/b)\,(i\omega_t) + 2^{1/2}\,||\boldsymbol{\omega}_x||^2\,(a/b)^2. \tag{35}$$

Thus, if we define

$$H(i\boldsymbol{\omega}_x, i\omega_t) = \frac{1}{(i\omega_t)^2 + 2^{5/4} \cos(\pi/8)\,||\boldsymbol{\omega}_x||\,(a/b)\,(i\omega_t) + 2^{1/2}\,||\boldsymbol{\omega}_x||^2\,(a/b)^2}. \tag{36}$$

then H is a time-stable transfer function such that

$$S(\boldsymbol{\omega}_x, \omega_t) \approx H(i\boldsymbol{\omega}_x, i\omega_t)\, S_w(\boldsymbol{\omega}_x)\, H(-i\boldsymbol{\omega}_x, -i\omega_t) \tag{37}$$

where

$$\begin{aligned} S_w(\boldsymbol{\omega}_x) &= \frac{2\sigma^2 \pi^{(d+1)/2}}{a\,||\boldsymbol{\omega}_x||\, b^{d-1}} \exp\left(-\frac{||\boldsymbol{\omega}_x||^2}{4b^2}\right) 2 \left(\frac{a^2 ||\boldsymbol{\omega}_x||^2}{b^2}\right)^2 \\ &= \left(\frac{4\sigma^2 \pi^{(d+1)/2} a^3}{b^{d+5}}\right) ||\boldsymbol{\omega}_x||^3 \exp\left(-\frac{||\boldsymbol{\omega}_x||^2}{4b^2}\right) \end{aligned} \tag{38}$$

Now let $w(\mathbf{x}, t)$ be a time-white Gaussian process with spectral density function $Q_w(\mathbf{x}) = \mathcal{F}_x^{-1}[S_w(\boldsymbol{\omega}_x)]$ and define the operators

$$\begin{aligned} \mathcal{A}_0 &= \mathcal{F}_x^{-1}[2^{1/2}\, ||\boldsymbol{\omega}_x||^2\, (a/b)^2] \\ \mathcal{A}_1 &= \mathcal{F}_x^{-1}[2^{5/4} \cos(\pi/8)\, ||\boldsymbol{\omega}_x||\, (a/b)], \end{aligned} \tag{39}$$

then the process $f(\mathbf{x}, t)$ approximately has the covariance function $C(\mathbf{x}, t)$:

$$\frac{\partial^2 f(\mathbf{x}, t)}{\partial t^2} + \mathcal{A}_1 \frac{\partial f(\mathbf{x}, t)}{\partial t} + \mathcal{A}_0 f(\mathbf{x}, t) = w(\mathbf{x}, t). \tag{40}$$

The first of the operators is just

$$\mathcal{A}_0 = 2^{1/2}\ (a/b)^2\, \mathcal{F}_x^{-1}[||\boldsymbol{\omega}_x||^2] = -2^{1/2}\ (a/b)^2\, \nabla^2 \tag{41}$$

The second operator can be written as

$$\mathcal{A}_1 = 2^{5/4} \cos(\pi/8)\, (a/b)\, \mathcal{F}_x^{-1}[||\boldsymbol{\omega}_x||] = 2^{5/4} \cos(\pi/8)\, (a/b) \sqrt{-\nabla^2} \tag{42}$$

In numerical computations the operator square root can be usually easily implemented. Thus the resulting pseudo-differential evolution equation is of the form

$$\frac{\partial}{\partial t} \begin{pmatrix} f(\mathbf{x}, t) \\ \frac{\partial f(\mathbf{x}, t)}{\partial t} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ c_0\, \nabla^2 & -c_1 \sqrt{-\nabla^2} \end{pmatrix} \begin{pmatrix} f(\mathbf{x}, t) \\ \frac{\partial f(\mathbf{x}, t)}{\partial t} \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} w(\mathbf{x}, t), \tag{43}$$

where $c_0 = 2^{1/2}\ (a/b)^2$ and $c_1 = 2^{5/4} \cos(\pi/8)\, (a/b)$ are constants.

To compute approximation to the covariance function with scalar x, let's approximate the operators with their Dirichlet counterparts on finite interval $[-L, L]$. Consider the eigenvalue problem

$$-\nabla^2 v_n(x) = -\frac{\partial^2 v_n(x)}{\partial x^2} = \lambda_n^2\, v_n(x), \qquad v_n(-L) = v_n(L) = 0, \tag{44}$$

The normalized (squared) eigenvalues and orthonormal eigenfunctions for $n = 1, 2, \ldots$ are:

$$\begin{aligned} \lambda_n &= \frac{n\,\pi}{2L} \\ v_n(x) &= \sqrt{\frac{1}{L}}\ \sin\left(\frac{n\,\pi\,(x + L)}{2L}\right). \end{aligned} \tag{45}$$

Thus the 1d Laplacian can be associated with the formal kernel

$$K_0(x,x') = -\sum_n \lambda_n^2 \, v_n(x) \, v_n(x'), \tag{46}$$

such that

$$\nabla^2 f(x,t) = \int K_0(x,x') \, f(x,t) \, dx \tag{47}$$

If we expand $f(x,t)$ on the basis $\{v_n(x)\}$ then we have

$$f(x,t) = \sum_n f_n(t) \, v_n(x). \tag{48}$$

where $f_n(t) = \int f(x,t) \, v_n(x) \, dx$. Thus

$$\begin{aligned} \nabla^2 f(x,t) &= \int K_0(x,x') \, f(x,t) \, dx \\ &= -\sum_{n,n'} \lambda_n^2 \, v_n(x) \, v_n(x') \, f_{n'}(t) \, v_{n'}(x) \, dx \\ &= -\sum_{n,n'} \lambda_n^2 \, v_n(x) \, \delta_{n,n'} \, f_{n'}(t) \\ &= -\sum_n \lambda_n^2 \, v_n(x) \, f_n(t). \end{aligned} \tag{49}$$

The square root operator $\sqrt{-\nabla^2}$ now has the formal kernel

$$K_1(x,x') = \sum_n \lambda_n \, v_n(x) \, v_n(x'). \tag{50}$$

We can now form (random) series expansion for $w(x,t)$ as follows:

$$\begin{aligned} w(x,t) &= \sum_n w_n(t) \, v_n(x) \\ w_n(t) &= \int w(x,t) \, v_n(x) \, dx. \end{aligned} \tag{51}$$

The differential equation can now be expressed in terms of the basis coefficients as follows:

$$\frac{d}{dt} \begin{pmatrix} f_n(t) \\ \frac{df_n(t)}{dt} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -c_0 \, \lambda_n^2 & -c_1 \, \lambda_n \end{pmatrix} \begin{pmatrix} f_n(t) \\ \frac{df_n(t)}{dt} \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} w_n(t). \tag{52}$$

which should be true for all n. The joint spectral density $\tilde{\mathbf{Q}}$ for the process noise can be derived as follows:

$$\begin{aligned} \mathrm{E}[w_n(t) \, w_m(s)] &= \mathrm{E}[\iint w(x,t) \, v_n(x) \, w(x',s) \, v_m(x') \, dx \, dx'] \\ &= \iint v_n(x) \; \mathrm{E}[w(x,t) \, w(x',s)] \, v_m(x') \, dx \, dx' \\ &= \iint v_n(x) \, Q_c(x-x') \, v_m(x') \, dx \, dx' \, \delta(t-s). \end{aligned} \tag{53}$$

i.e.,

$$\tilde{Q}_{nm} = \iint v_n(x)\, \mathbf{L}\, Q_c(x - x')\, \mathbf{L}^T\, v_m(x')\, dx\, dx'. \tag{54}$$

with $\mathbf{L} = (0, 1)$. Thus we have a model of the form

$$d\mathbf{f} = \mathbf{A}\, \mathbf{f}\, dt + d\mathbf{W}, \tag{55}$$

where $\mathbf{f} = (f_1, df_1/dt, f_2, df_2/dt, \ldots,)$ and

$$\mathbf{A} = \begin{pmatrix} \begin{pmatrix} 0 & 1 \\ -c_0\, \lambda_1^2 & -c_1\, \lambda_1 \end{pmatrix} & & \\ & \begin{pmatrix} 0 & 1 \\ -c_0\, \lambda_2^2 & -c_1\, \lambda_2 \end{pmatrix} & \\ & & \ddots \end{pmatrix} \tag{56}$$

and the diffusion matrix of $\mathbf{W}$ is $\tilde{\mathbf{Q}}$. The measurement model is then

$$y_k = \tilde{\mathbf{H}}_k\, \mathbf{f} + e_k, \tag{57}$$

where $\tilde{\mathbf{H}}_k = (v_1(x_k)\ 0\ v_2(x_k)\ 0\ \cdots)$.

The equation for the mean $\mathbf{m}$ and covariance $\mathbf{P}$ of $\mathbf{f}$ are now given as

$$\frac{d\mathbf{m}}{dt} = \mathbf{A}\, \mathbf{m} \tag{58}$$

$$\frac{d\mathbf{P}}{dt} = \mathbf{A}\, \mathbf{P} + \mathbf{P}\, \mathbf{A}^T + \tilde{\mathbf{Q}}. \tag{59}$$

Let $\mathbf{P}_\infty$ be the solution to the equation

$$\mathbf{A}\, \mathbf{P}_\infty + \mathbf{P}_\infty\, \mathbf{A}^T + \tilde{\mathbf{Q}} = 0 \tag{60}$$

Then we have

$$\mathbf{C}_{\mathbf{f}}(\tau) = \mathrm{E}[\mathbf{f}(t)\, \mathbf{f}^T(t + \tau)] = \begin{cases} \mathbf{P}_\infty\, \exp(\tau\, \mathbf{A})^T & , \quad \text{for } \tau \geq 0 \\ \exp(-\tau\, \mathbf{A})\, \mathbf{P}_\infty & , \quad \text{for } \tau < 0 \end{cases} \tag{61}$$

where

$$\exp(\tau\, \mathbf{A}) = \begin{pmatrix} \exp\left\{ \begin{pmatrix} 0 & 1 \\ -c_0\, \lambda_1^2 & -c_1\, \lambda_1 \end{pmatrix} \tau \right\} & & \\ & \exp\left\{ \begin{pmatrix} 0 & 1 \\ -c_0\, \lambda_2^2 & -c_1\, \lambda_2 \end{pmatrix} \tau \right\} & \\ & & \ddots \end{pmatrix} \tag{62}$$

If we define $\mathbf{v}(x) = (v_1(x), v_2(x), \ldots)$, then we have

$$f(x, t) = \sum_n f_n(t)\, v_n(x) = \mathbf{v}^T(x)\, \mathbf{H}\, \mathbf{f}(t) \tag{63}$$

where $\mathbf{H}$ is a matrix with elements $H_{j,2j} = 1$ and thus

$$\begin{aligned}
\mathrm{E}[f(x,t)\, f(x+\xi, t+\tau)] &= \mathrm{E}[\mathbf{v}^T(x)\, \mathbf{H}\, \mathbf{f}(t)\, \mathbf{v}^T(x+\xi)\, \mathbf{H}\, \mathbf{f}(t+\tau)] \\
&= \mathbf{v}^T(x)\, \mathbf{H}\ \mathrm{E}[\mathbf{f}(t)\, \mathbf{f}(t+\tau)]\, \mathbf{H}^T\ \mathbf{v}(x+\xi) \\
&= \mathbf{v}^T(x)\, \mathbf{H}\, \mathbf{C}_{\mathbf{f}}(\tau) \mathbf{H}^T\ \mathbf{v}(x+\xi).
\end{aligned} \tag{64}$$

Thus we can approximate the covariance function defined by the stochastic equation by

$$C_f(x,t) \approx \mathbf{v}^T(0)\, \mathbf{H}\, \mathbf{C}_{\mathbf{f}}(t)\, \mathbf{H}^T\ \mathbf{v}(x). \tag{65}$$

The covariance function can be now numerically computed by using a finite number of terms from this expansion. The Kalman filtering and RTS smoothing based estimation solution can be done by using a finite number of series terms in dynamic model (55) and measurement model (57).

4 Details of Modeling US Monthly Precipitation and Temperature Data

4.1 Model

We implemented the separable spatio-temporal GPs as finite-dimensional SDEs of form as

$$d\mathbf{f}(t) = \mathbf{A}\, \mathbf{f}(t)\, dt + \mathbf{L}\, d\mathbf{W}(t), \tag{66}$$

where matrix $\mathbf{A}$ is a $dN \times dN$ block diagonal matrix, where the $N \times N$ blocks are constructed in such a way that they determine the desired temporal covariance function $C_t(t)$ for the n components (see Hartikainen and Särkkä, 2010, for more details). In this example we used the Matérn temporal covariance model. For the spatial covariance $C_x(\mathbf{x})$ we used 2-dimensional Matérn covariance ($\nu = 3/2$), which is used in forming the elements of diffusion matrix $\mathbf{Q}_c$ of $\mathbf{W}(t)$.

To further lighten up the computations we formed the finite-dimensional model (66) to a latent *inducing process* $\mathbf{u}(t)$ on fixed spatial locations $\{\mathbf{x}_u^i\}_{i=1}^m$, and constructed a linear-Gaussian mapping from the inducing process to a infinite-dimensional latent process as $f(\mathbf{x},t)|\mathbf{u}(t) \sim N(\mathbf{H}(\mathbf{x})\mathbf{u}(t), \mathbf{R}(\mathbf{x})$, where matrices in the mapping are set to $\mathbf{H}(\mathbf{x}) = \mathbf{C}_{\mathbf{f},\mathbf{u}}\mathbf{C}_{\mathbf{u},\mathbf{u}}^{-1}$ and $\mathbf{R}(\mathbf{x}) = \mathrm{diag}(\mathbf{C}_{\mathbf{f},\mathbf{f}} - \mathbf{C}_{\mathbf{x},\mathbf{u}}\, \mathbf{C}_{\mathbf{u},\mathbf{u}}^{-1}\, \mathbf{C}_{\mathbf{u},\mathbf{x}})$, where the covariance terms are evaluated with the spatial covariance function C_x. This can be seen as dynamic formulation of *fully independent conditional* (FIC) sparse approximation recently proposed in the standard GP regression framework. Different approximations can be constructed by choosing the matrices $\mathbf{H}$ and $\mathbf{R}$ appropriately.

To achieve the computational efficiency (i.e., $\mathcal{O}(dm^2)$ complexity in measurement updates) with the low-rank model one can use the matrix inversion lemma to avoid the inversion of $n \times n$ matrix and rather invert a $m \times m$ matrix. In Kalman filtering context the matrix inversion lemma is commonly implemented such that the estimated states and covariances are replaced with *information vectors* and *information matrices*, which are defined as $\mathbf{I}_k = \mathbf{P}_k^{-1}$ and

$\mathbf{i}_k = \mathbf{P}_k^{-1}\mathbf{m}_k$. This formulation is Kalman filter is commonly termed as *information filter* (Grewal and Andrews, 2001). In addition to computational efficiency the information filter is more numerically robust with the low-rank model, which is particularly important in marginal likelihood based hyperparameter learning.

4.2 Data

The data we consider in the paper consists of monthly precipitation and temperature minimum/maximum measurements [1] collected in the US from years 1895-1997. There are 11918 measurements stations for the precipitation data and 8125 for the temperatures. Subsets of this data were used by Paciorek and Schervish (2006) and Vanhatalo and Vehtari (2008) to assess spatial regression models. High fraction of the measurements is missing, and our aim is to fill out the missing measurements by taking account of the spatio-temporal correlations in the data. As the size of original data is very large we focus on (roughly) the same subset of data as in Paciorek and Schervish (2006). The subset is collected from a rectangular area ($[-109.5, -101] \times [36.5, 41.5]$ lon/lat) around Colorado and comprises of 502 stations for the precipitation and 423 for the temperature readings. The total number of measurements in the subset are 372873 for precipitation, 336156 for maximum temperature and 336720 for minimum temperature.

Locations of the measurements stations for precipitation data are shown in Figure 1. Examples of time-series of each data set are shown in Figure 1. The time dynamics of precipitation are much more chaotic than the naturally periodic behavior of temperature readings.

Figure 1: Locations of the measurement stations in the precipitation data. Black dots represent the locations in the whole data, and red dots the locations in the subsample, which used in the experiments. Plots with temperature data are similar, but the number of stations is smaller.

[1] http://www.image.ucar.edu/GSP/Data/US.monthly.met/

References

Cressie, N. and Huang, H.-C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *JASA*, 94(448):1330–1340.

Grewal, M. S. and Andrews, A. P. (2001). *Kalman Filtering, Theory and Practice Using MATLAB*. Wiley Interscience.

Hartikainen, J. and Särkkä, S. (2010). Kalman filtering and smoothing solutions to temporal Gaussian process regression models. In *Proceedings of MLSP*.

Jazwinski, A. (1970). *Stochastic Processes and Filtering Theory*. Academic Press.

Karatzas, I. and Shreve, S. E. (1991). *Brownian Motion and Stochastic Calculus*. Springer.

Øksendal, B. (2003). *Stochastic Differential Equations: An Introduction with Applications*. Springer, 6th edition.

Paciorek, C. and Schervish, M. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, 17(5):483–506.

Råde, L. and Westergren, B. (2004). *Mathematics Handbook*. Studentlitteratur, 5th edition.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Vanhatalo, J. and Vehtari, A. (2008). Modelling local and global phenomena with sparse Gaussian processes. In *Proceedings of UAI*.

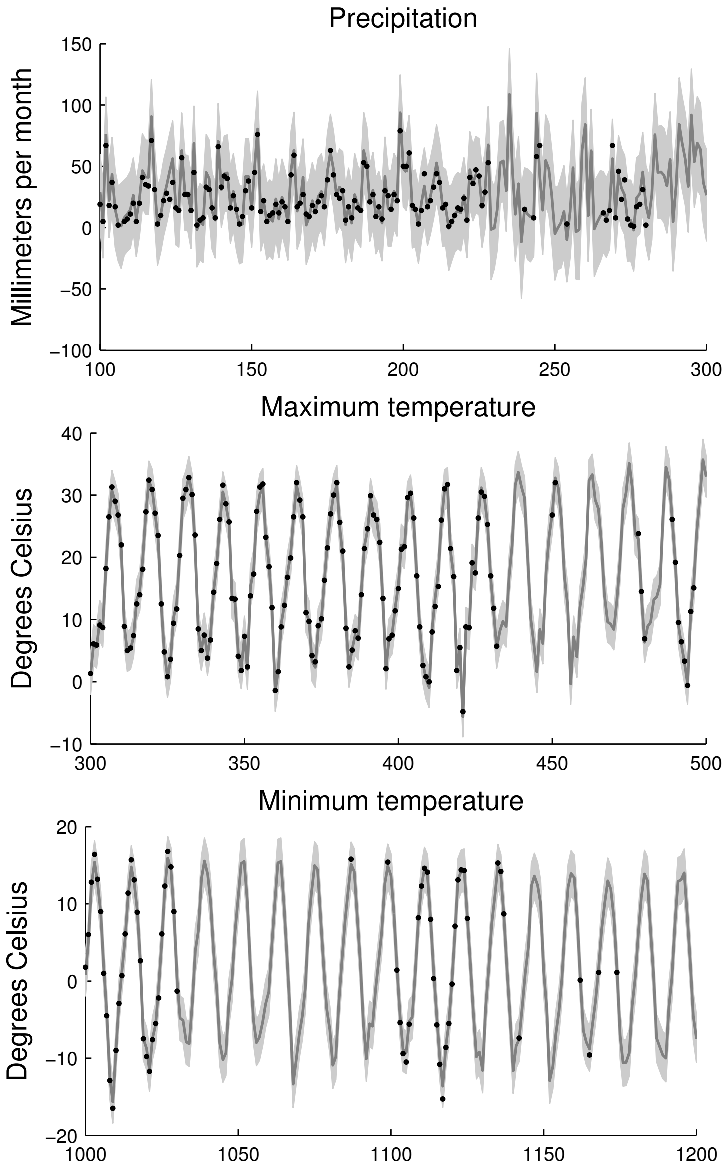

Figure 2: Example time series of each data and estimate of them obtained with STGP ($\nu = 3/2$). Black dots are the measurements, dark gray the mean estimate and light gray the 95% uncertainty.