Sparse Spatio-Temporal Gaussian Processes with General Likelihoods

Jouni Hartikainen, Jaakko Riihimäki and Simo Särkkä

Dept. of Biomedical Engineering and Computational Science
Aalto University, Finland
jmjharti@cc.hut.fi,jaakko.riihimaki@tkk.fi,simo.sarkka@tkk.fi

Abstract. In this paper, we consider learning of spatio-temporal processes by formulating a Gaussian process model as a solution to an evolution type stochastic partial differential equation. Our approach is based on converting the stochastic infinite-dimensional differential equation into a finite dimensional linear time invariant (LTI) stochastic differential equation (SDE) by discretizing the process spatially. The LTI SDE is time-discretized analytically, resulting in a state space model with linear-Gaussian dynamics. We use expectation propagation to perform approximate inference on non-Gaussian data, and show how to incorporate sparse approximations to further reduce the computational complexity. We briefly illustrate the proposed methodology with a simulation study and with a real world modelling problem.

Keywords: Gaussian processes, spatio-temporal data, expectation propagation, sparse approximations

1 Introduction

Over the last decades Gaussian process (GP) based methods [1] have steadily increased popularity as prominent tools for data analysis in several fields, including spatial statistics, epidemiology and machine learning. Although, in the common machine learning setting the modeled phenomena are assumed to be static in time, learning of time dependent spatio-temporal models have recently gained much interest. So far, the application of generic GP techniques to spatio-temporal data has been hindered by the steep increase in computational requirements with respect to the number of data points.

In this article, we show how evolution type stochastic partial differential equations [2] can be used as flexible prior models in spatio-temporal learning. In our approach, the Gaussian spatio-temporal prior processes are modeled as linear time-invariant stochastic partial differential equations, and the measurement models are assumed to be generic conditional distribution models for the measurements. Formulating the model this way enables us to make use of the Markov property inherent in the system to perform inference sequentially. Furthermore, we show how to incorporate the recently proposed sparse GP approximations [3, 4] into the spatio-temporal formulation, which further reduces the computational

burden. When combined with expectation propagation (EP) [5] approximate inference scheme the computations are very cheap, enabling accurate inference on large-scale spatio-temporal data sets.

As such, learning of spatio-temporal systems which are modeled as stochastic differential equations is a mature subject and has been much studied in control engineering under the names distributed parameter systems [6] and infinite-dimensional (Kalman) filtering [7]. More recently, the Bayesian Kalman filtering approach to spatio-temporal estimation has been studied, for example, in geostatistics [8–10] as well as in statistical inversion theory [11, 12]. In machine learning context the usage of differential equations and partial differential equations for encoding prior information into Gaussian process regression models has recently been discussed in [13].

2 Model and Methods

2.1 Spatio-Temporal Gaussian Processes

In this paper we consider evolution type stochastic partial differential equations (SPDEs) [2] of the following form:

$$\frac{\partial \mathbf{x}(t,\mathbf{r})}{\partial t} = \mathcal{A}_r \mathbf{x}(t,\mathbf{r}) + \mathcal{L}_r \mathbf{w}(t,\mathbf{r}), \quad \mathbf{y}_k \sim \prod_{i=1}^{n} p(y_{ki} \,|\, \mathbf{x}(t_k,\mathbf{r}_i)), \tag{1}$$

where $\mathbf{x}(t,\mathbf{r})$ denotes the latent spatio-temporal prior Gaussian process depending on the time $t \geq 0$ and spatial location $\mathbf{r} \in \mathbb{D}$ on some bounded domain $\mathbb{D} \subset \Re^d$, and $\mathbf{y}_k = (y_{k1}, \ldots, y_{kn})$ are the measurements. $\mathcal{A}_r$ and $\mathcal{L}_r$ are linear operators acting on the variable $\mathbf{r}$. The noise process $\mathbf{w}(t,\mathbf{r})$ is a Gaussian process with $\mathbf{r}$-dimensional covariance function of the time-white form $k(t,\mathbf{r};t',\mathbf{r}') = \delta(t-t')\,k(\mathbf{r},\mathbf{r}')$, where $k(\mathbf{r},\mathbf{r}')$ is some suitably chosen spatial covariance function. Since $\mathcal{A}_r$ and $\mathcal{L}_r$ are linear operators and $\mathbf{w}(t,\mathbf{r})$ is a Gaussian process, $\mathbf{x}(t,\mathbf{r})$ is also a Gaussian process.

Often in Bayesian inference for Gaussian processes the model is formulated in terms of time-space covariance function $k(t,\mathbf{r};t',\mathbf{r}')$ instead of a SPDE. However, as shown in [14] there is one-to-one mapping between a large class of temporal covariance functions (including the Matérn class) and linear state space models. Similarly, there is an analogous one-to-one mapping between spatio-temporal covariance functions and SPDEs. In the case of separable covariance functions of the form $k(t,\mathbf{r};t',\mathbf{r}') = k_t(t,t')\,k_s(\mathbf{r},\mathbf{r}')$ where k_t and k_s are appropriate temporal and spatial covariance functions, the mapping becomes particularly simple and computationally efficient. In our examples we shall consider models of this form.

After obtaining a set of observations $\mathbf{y}_{1:T} = \{\mathbf{y}, \ldots, \mathbf{y}_T\}$ the aim is to infer the state posterior distribution $p(\mathbf{x}(t,\mathbf{r})|\mathbf{y}_{1:T})$. In practice, $\mathbf{x}(t,\mathbf{r})$ is discretized with respect to space and time to make the model tractable. Additionally, the dynamic model typically has few hyperparameters $\theta = (\theta_1, \ldots, \theta_p)$, which need to be learned. These can include, for instance, the spatial length scales and magnitudes of the noise process $\mathbf{w}(t,\mathbf{r})$ as well as possible parameters of the operators $\mathcal{A}_r$ and $\mathcal{L}_r$.

2.2 Making the Model Tractable

A simple way to convert a stochastic partial differential equation model into tractable form is to use discretization. For example, by using a finite difference or finite basis type of approximation in the spatial dimension, the infinite-dimensional SPDE model can be transformed into finite-dimensional SDE:

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

where matrices $\mathbf{F}$ and $\mathbf{L}$ are finite dimensional approximations to the linear operators $\mathcal{A}_r$ and $\mathcal{L}_r$, and $\mathbf{x}(t) = (\mathbf{x}_1(t), \ldots, \mathbf{x}_n(t))$ is the state of the process at a finite set of spatial points $\{\mathbf{r}_1, \ldots, \mathbf{r}_n\}$. GPs with separable covariance functions result also in models of this form, where $\mathbf{n}(t)$ has the covariance function $\delta(t - t')\, k_s(\mathbf{r}, \mathbf{r}')$ and $\mathbf{F}$ is a $hn \times hn$ block diagonal matrix, where the $h \times h$ blocks are constructed in such a way that they determine the desired temporal covariance function $k_t(t, t')$ for the n components (see [14] for more details).

In practice, we are interested in the values of the Gaussian process at discrete points of time, say, $t \in \{t_1, t_2, \ldots\}$. By using the well known methods from linear systems theory [15], the continuous time LTI model above can be transformed into discrete time model of the following form:

$$\mathbf{x}_k = \mathbf{A}_{k-1}\,\mathbf{x}_{k-1} + \mathbf{q}_{k-1}, \quad \mathbf{q}_{k-1} \sim N(\mathbf{0}, \mathbf{Q}_{k-1}), \quad \mathbf{y}_k \sim p(\mathbf{y}_k \,|\, \mathbf{x}_k), \tag{3}$$

where the matrices $\mathbf{A}_{k-1}$ and $\mathbf{Q}_{k-1}$ have analytic solutions (see, e.g., [15]).

2.3 Sparse Approximations

Suppose that we have a GP prior on n latent variables $\mathbf{x} \in \Re^n$ with input features $\{\mathbf{r}_x^i\}_{i=1}^n$ as $\mathbf{x} \sim N(0, \mathbf{K}_{\mathbf{x},\mathbf{x}})$. The problem of this approach is the $\mathcal{O}(n^3)$ scaling of computations in the inference. The recently developed sparse approximations [3, 4] are aimed to mitigate these problems by placing a GP prior on a smaller set of m *inducing variables* $\mathbf{u} \in \Re^m$ (with own input features $\{\mathbf{r}_u^i\}_{i=1}^m$) as $\mathbf{u} \sim N(0, \mathbf{K}_{\mathbf{u},\mathbf{u}})$, and then setting a linear-Gaussian relationship between the inducing variables $\mathbf{u}$ and the actual latent variables $\mathbf{x}$ as $\mathbf{x}|\mathbf{u} \sim N(\mathbf{H}\,\mathbf{u}, \mathbf{R})$. Different approximations can be constructed by choosing the matrices $\mathbf{H}$ and $\mathbf{R}$ appropriately. For example, by choosing $\mathbf{H} = \mathbf{K}_{\mathbf{x},\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}$ and $\mathbf{R} = \mathrm{diag}(\mathbf{K}_{\mathbf{x},\mathbf{x}} - \mathbf{K}_{\mathbf{x},\mathbf{u}}\,\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\,\mathbf{K}_{\mathbf{u},\mathbf{x}})$ we obtain the *fully independent conditional* (FIC) approximation, which we use as an example during the rest of this paper. Due to linear-Gaussian formulation, the values of $\mathbf{u}$ can always be integrated out analytically during the inference[1], and by using the well-known matrix inverse lemma the computations can be significantly reduced if $\mathbf{R}$ is of such form that it can be inverted easily. For example, if $\mathbf{R}$ is diagonal the complexity is $\mathcal{O}(nm^2)$.

To translate these ideas to spatio-temporal models we propose to formulate a separable spatio-temporal GP prior model for *inducing process* $\mathbf{u}(t) \in \Re^m$ as

$$\frac{d\mathbf{u}(t)}{dt} = \mathbf{F}_u\mathbf{u}(t) + \mathbf{L}_u\mathbf{n}(t), \tag{4}$$

[1] The input features of $\mathbf{u}$, however, have an impact on the result.

and the observation model as

$$\mathbf{x}_k|\mathbf{u}_k \sim N(\mathbf{H}_k\mathbf{u}_k, \mathbf{R}_k),\ \mathbf{y}_k \sim p(\mathbf{y}_k|\mathbf{x}_k). \tag{5}$$

This formulation allows also to specify more general models by defining $\mathbf{H}_k$ and $\mathbf{R}_k$ appropriately. For example, we can formulate additive models, in which there are separate spatial, temporal and spatio-temporal components as well as covariates, which have linear or fixed basis effects. This approach allows also to predict the process on arbitrary input $\mathbf{r}$ since we can write the conditional as $\mathbf{x}(t,\mathbf{r})|\mathbf{u}(t) \sim N(\mathbf{H}(\mathbf{r})\mathbf{u}(t), \mathbf{R}(\mathbf{r}))$, which we can easily integrate over the posterior of $\mathbf{u}(t)$ to get the marginal of $\mathbf{x}(t,\mathbf{r})$.

2.4 Expectation Propagation for Dynamic Systems

With generic GPs and non-Gaussian likelihoods expectation propagation (EP) [5] has been shown to give state-of-the-art performance compared to other deterministic inference methods [16]. For dynamic systems EP was first introduced by [17] and later extended for non-linear/Gaussian [18] and non-linear/Poisson smoothing problems [19]. With EP, Gaussian approximations are made only in the state space, avoiding possible difficulties arising with the Kalman filtering type of methods [15]. In this article we apply EP to spatio-temporal GPs with non-Gaussian likelihoods.

The central idea of EP is to factor the smoothing distribution as

$$p(\mathbf{x}_{1:T}|\mathbf{y}_{1:T}) \approx \hat{p}(\mathbf{x}_{1:T}) \propto \prod_{k=1}^{T} \alpha_k(\mathbf{x}_k)\beta_k(\mathbf{x}_k), \tag{6}$$

where the forward and backward messages $\alpha_k(\mathbf{x}_k) \propto p(\mathbf{x}_k|\mathbf{y}_{1:k})$ and $\beta_k(\mathbf{x}_k) \propto p(\mathbf{y}_{k+1:T}|\mathbf{x}_k, \mathbf{y}_{1:k})$ are iteratively refined such that the Kullback-Leibler (KL) divergence from the true posterior $p(\mathbf{x}_{1:T}|\mathbf{y}_{1:T})$ to an approximation $\hat{p}(\mathbf{x}_{1:T})$ is minimized. While the global minimization is intractable, in EP the minimization is performed by sequentially minimizing the KL divergence from $p(\mathbf{x}_{k-1}, \mathbf{x}_k|\mathbf{y}_{1:T}) \propto \alpha_{k-1}(\mathbf{x}_{k-1})p(\mathbf{x}_k|\mathbf{x}_{k-1})p(\mathbf{y}_k|\mathbf{x}_k)\beta_k(\mathbf{x}_k)$ to an approximation $\hat{p}(\mathbf{x}_{k-1}, \mathbf{x}_k)$. The messages $\alpha_k(\mathbf{x}_k)$ and $\beta_k(\mathbf{x}_k)$ are typically chosen to be members of exponential family (in our case un-normalized Gaussians), and such cases the minimization of KL divergence is equivalent to moment matching. In our case this means that the approximation $\hat{p}(\mathbf{x}_{k-1}, \mathbf{x}_k)$ is Gaussian, and in next section we briefly detail how to seek its moments efficiently for the class of models considered here. After obtaining $\hat{p}(\mathbf{x}_{k-1}, \mathbf{x}_k)$, the messages are updated in forward pass as $\alpha_k^{\text{new}}(\mathbf{x}_k) = \int \hat{p}(\mathbf{x}_{k-1}, \mathbf{x}_k)d\mathbf{x}_{k-1}/\beta_k(\mathbf{x}_k)$ and in backward pass as $\beta_{k-1}^{\text{new}}(\mathbf{x}_{k-1}) = \int \hat{p}(\mathbf{x}_{k-1}, \mathbf{x}_k)d\mathbf{x}_k/\alpha_{k-1}(\mathbf{x}_{k-1})$. Usually several forward and backward passes over the data are needed to achieve convergence.

Approximating the Two-Slice Posterior We now seek to find a Gaussian approximation for $p(\mathbf{x}_{k-1}, \mathbf{x}_k|\mathbf{y}_{1:T})$ via moment matching. First, the product of densities $p_*(\mathbf{x}_{k-1}, \mathbf{x}_k) = \alpha_{k-1}(\mathbf{x}_{k-1})p(\mathbf{x}_k|\mathbf{x}_{k-1})$ can be written as

$$p_*(\mathbf{x}_{k-1}, \mathbf{x}_k) \propto N(\mathbf{x}_{k-1,k}|\mathbf{m}^*_{k-1,k}, \mathbf{P}^*_{k-1,k}), \tag{7}$$

where

$$\mathbf{m}^*_{k-1,k} = \begin{bmatrix} \mathbf{m}^\alpha_{k-1} \\ \mathbf{m}^*_k \end{bmatrix}, \;\; \mathbf{P}^*_{k-1,k} = \begin{bmatrix} \mathbf{P}^\alpha_{k-1} & \mathbf{D}^T_k \\ \mathbf{D}_k & \mathbf{P}^*_k \end{bmatrix} \tag{8}$$

and

$$\mathbf{m}^*_k = \mathbf{A}_{k-1}\mathbf{m}^\alpha_{k-1}, \;\; \mathbf{D}_k = \mathbf{A}_{k-1}\mathbf{P}^\alpha_{k-1}, \;\; \mathbf{P}^*_k = \mathbf{A}_{k-1}\mathbf{P}^\alpha_{k-1}\mathbf{A}^T_{k-1} + \mathbf{Q}_{k-1}. \tag{9}$$

This can also be decomposed as $p_*(\mathbf{x}_{k-1}, \mathbf{x}_k) \propto p_*(\mathbf{x}_k)p_*(\mathbf{x}_{k-1}|\mathbf{x}_k)$, where

$$\begin{aligned} &p_*(\mathbf{x}_k) = N(\mathbf{x}_k|, \mathbf{m}^*_k, \mathbf{P}^*_k), \qquad p_*(\mathbf{x}_{k-1}|\mathbf{x}_k) = N(\mathbf{x}_{k-1}|\mathbf{m}^*_{k-1|k}, \mathbf{P}^*_{k-1|k}), \\ &\mathbf{m}^*_{k-1|k} = \mathbf{m}^\alpha_{k-1} + \mathbf{D}^T_k[\mathbf{P}^*_k]^{-1}(\mathbf{x}_k - \mathbf{m}^*_k), \;\; \mathbf{P}^*_{k-1|k} = \mathbf{P}^\alpha_{k-1} - \mathbf{D}^T_k[\mathbf{P}^*_k]^{-1}\mathbf{D}_k. \end{aligned} \tag{10}$$

The backward message can be incorporated by simply using the product rule of Gaussian distribution to get $p_{**}(\mathbf{x}_k) = p_*(\mathbf{x}_k)\beta_k(\mathbf{x}_k) \propto N(\mathbf{x}_k|\mathbf{m}^{**}_k, \mathbf{P}^{**}_k)$.

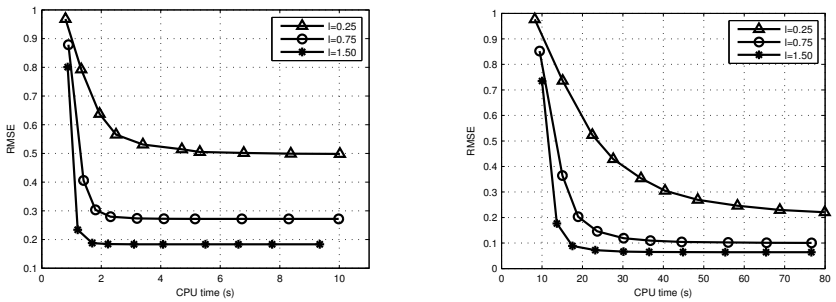
The posterior is now of form $\hat{p}(\mathbf{x}_{k-1}, \mathbf{x}_k) \propto p_*(\mathbf{x}_{k-1}|\mathbf{x}_k)p_{**}(\mathbf{x}_k)p(\mathbf{y}_k|\mathbf{x}_k)$. By using the Bayes' rule we can write $\hat{p}(\mathbf{x}_k) \propto p_{**}(\mathbf{x}_k)p(\mathbf{y}_k|\mathbf{x}_k)$ when we treat $p_{**}(\mathbf{x}_k)$ as a prior for $\mathbf{x}_k$. Generally this is not of an analytically tractable form, but we can seek Gaussian approximations by applying any approximate inference scheme applicable to GPs with non-Gaussian likelihoods. Common approaches are Laplace approximation or EP (see, e.g.,[1]). If we use sparse approximations or other generalized observation models, the dynamic model would be defined for $\mathbf{u}_k$ and the prior for the "moment matching" algorithm is $p_{**}(\mathbf{x}_k) \propto N(\mathbf{H}_k\,\mathbf{m}^{**}_k, \mathbf{H}_k\,\mathbf{P}^{**}_k\,\mathbf{H}^T_k + \mathbf{R}_k)$. Since the covariance of this prior is of same form as in sparse GPs, we can use same tricks as presented, e.g., in [20] to speed up the inference. With this we achieve the overall complexity $\mathcal{O}(NTnm^2)$, where N is the number of EP iterations across the time sequence (in our examples we used $N = 3$, which we empirically observed to be sufficient).

After obtaining an approximation $\hat{p}(\mathbf{x}_k) \propto N(\mathbf{x}_k|\mathbf{m}_k, \mathbf{P}_k)$, the (marginalized) posterior of $\mathbf{x}_{k-1}$ used in updating the backward messages can be obtained by combining $\hat{p}(\mathbf{x}_k)$ with (10), which results in Kalman smoothing like equations that are not stated here due to lack of space.

3 Results

We briefly show how to analyze log-Gaussian Cox process models by using the presented modelling framework. We consider two large sized examples: a simulation study highlighting the properties of our approach, and a real-world example concerning tropical rainforest point process data modelled recently by [21, 22].

The log-Gaussian Cox process can be formulated in practice such that the observations y_i in the region w_i are Poisson distributed with mean $|w_i| \exp(\eta(t_i, \mathbf{r}_i))$, where $|w_i|$ is the area of the subregion (in our examples constant), and $\eta(t, \mathbf{r})$ is the latent intensity field, which is given a spatial or spatio-temporal prior.



(a) $n \times T = 100 \times 100$ (b) $n \times T = 500 \times 500$

Fig. 1. Simulation study: Comparison of RMSE values versus the used CPU time for the considered data set sizes averaged over several simulation runs. The software was implemented on Matlab and ran on AMD Phenom II 3.5 GHz, 4GB RAM PC.

3.1 The Effect of Sparse Approximations

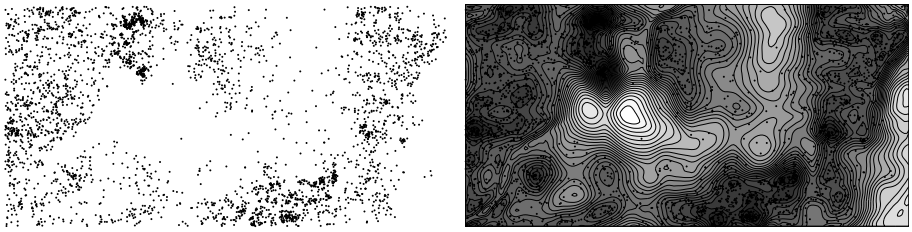
First we shall test how the sparse approximations affect the accuracy of the posterior estimate. For simplicity we consider here only two dimensional fields such that we treat one coordinate as time and the other as space. We simulate intensity fields η with a GP prior having a separable covariance function $k(t, r; t', r') = k_t(t, t')\, k_s(r, r')$, where both k_t and k_s are Matérn covariance functions with smoothness and magnitude parameters set to $\nu = 3/2$ and $\sigma^2 = 1$. We generate three different cases, in which the length scale parameter (common for both covariance functions) has the values $l \in \{0.25, 0.75, 1.5\}$. We generate data sets of size 100×100 and 500×500, and generate Poisson observations after generating the intensities. Given the observed data, we set the field to have a sparse GP prior and use EP to estimate its posterior. Figure 1 shows the RMSE values plotted against the used CPU time in cases of using different number of inducing variables between 2 and 70. It can be seen that the smoother the field the less number of inducing variables is needed to achieve accurate results.

3.2 Tropical Rainforest Data

We consider tropical rainforest data shown in Panel (a) of Figure 2. The data consists of 3605 trees in a rectangular rainforest area discretized into a 201×101 regular lattice. In each subregion also altitude and norm of the gradient are observed. Similarly as in [21, 22], we model the log of the mean parameter in Poisson distribution as

$$\eta_{ij} = \beta_0 + \beta_{\text{alt}}\, \text{alt}_{ij} + \beta_{\text{grad}}\, \text{grad}_{ij} + \mathbf{x}_{ij} + \epsilon_{ij}, \tag{11}$$

where β_0 is a base line effect, β_{alt} and β_{grad} the effects of the elevation and gradient values, $\mathbf{x}_{ij}$ a spatially structured effect and ϵ_{ij} a non-structure random effect. We place a sequential sparse GP prior for $\mathbf{x}_{ij}$ similarly as in previous



(a) Tree locations (b) Posterior mean of $\eta_{ij} - \epsilon_{ij}$

Fig. 2. Tropical rainforest data: (a) Data and (b) the mean estimate of $\eta_{ij} - \epsilon_{ij}$ produced by EP. We used Laplace's method with FIC ($m = 60$) in approximating the one-slice posteriors. The horizontal axis was treated as time and the vertical as space.

section and model the random effect as $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$. The mean estimate of the log intensity produced by EP is shown in Panel (b) of Figure 2. Hyperparameters of the model were optimized w.r.t (approximate) marginal likelihood $p(\mathbf{y}_{1:T}) = \prod_{k:1}^{T} p(\mathbf{y}_k|\mathbf{y}_{1:k-1})$. Although we could use the full spatio-temporal GP prior for the data considered here, by using FIC the computations were significantly faster (the optimization taking only few minutes of CPU time) without affecting result.

4 Conclusions

In this article we have shown how spatio-temporal Gaussian processes can be formed as linear-Gaussian state-space models that can be efficiently inferred by using sequential algorithms. We have shown the key details on how to implement EP for this class of GP priors with non-Gaussian observations. Moreover, we have shown how to incorporate the sparse approximations for further speeding up the computations. In future work we shall study wider class of spatio-temporal Gaussian processes with more general covariance functions and linear operators, implement a finite basis type of approximation to the SPDE by using the sparse approximations, marginalize over the hyperparameters numerically as in [21, 22] and apply the developed modelling framework to high dimensional data sets.

Acknowledgments. The authors would like to thank Finnish Doctoral Programme in Computational Sciences (FICS), Centre of Excellence in Computational Complex Systems Research (COSY) and Academy of Finland for financial support, and Aki Vehtari, Pasi Jylänki, Janne Ojanen and Jarno Vanhatalo for helpful discussions during the work.

References

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

2. Da Prato, G., Zabczyk, J.: Stochastic Equations in Infinite Dimensions. Cambridge University Press (1992)
3. Quinonero-Candela, J., Rasmussen, C.E.: A unifying view of sparse approximate Gaussian process regression. Journal Of Machine Learning Research 6(Oct), 1939–1959 (2005)
4. Snelson, E., Ghahramani, Z.: Sparse Gaussian process using pseudo-inputs. In Weiss, Y., Schlkopf, B., Platt, J., eds.: Advances in Neural Information Processing Systems. Volume 18. The MIT Press (2006)
5. Minka, T.: A family of algorithms for approximate Bayesian inference. PhD thesis, Massachusetts Institute of Technology (2001)
6. Ray, W.H., Lainiotis, D.G.: Distributed Parameter Systems. Dekker (1978)
7. Curtain, R.: A survey of infinite-dimensional filtering. SIAM Review 17(3), 395–411 (1975)
8. Wikle, C.K., Cressie, N.: A dimension-reduced approach to space-time Kalman filtering. Biometrika 86(4), 815–829 (1999)
9. Cressie, N., Wikle, C.K.: Space-time Kalman filter. In El-Shaarawi, A.H., Piegorsch, W.W., eds.: Encyclopedia of Environmetrics. Volume 4. John Wiley & Sons, Ltd, Chichester, 2045–2049 (2002)
10. Gelfand, A.E., Diggle, P.J., Fuentes, M., Guttorp, P.: Handbook of Spatial Statistics. Chapman & Hall/CRC (2010)
11. Kaipio, J., Somersalo, E.: Statistical and Computational Inverse Problems. Number 160 in Applied mathematical Sciences. Springer (2005)
12. Hiltunen, P., Särkkä, S., Nissilä, I., Lajunen, A., Lampinen, J.: State space regularization in the nonstationary inverse problem for diffuse optical tomography. Inverse Problems 27(2) (2011)
13. Alvarez, M., Lawrence, N.D.: Latent force models. In van Dyk, D., Welling, M., eds.: Proceedings of the Twelfth International Workshop on Artificial Intelligence and Statistics. 9–16 (2009)
14. Hartikainen, J., Särkkä, S.: Kalman Filtering and Smoothing Solutions to Temporal Gaussian Process Regression Models. In: Proceedings of IEEE International Workshop on Machine Learning for Signal Processing (MLSP). 379–384 (2010)
15. Bar-Shalom, Y., Li, X.R., Kirubarajan, T.: Estimation with Applications to Tracking and Navigation. Wiley Interscience (2001)
16. Nickisch, H., Rasmussen, C.: Approximations for binary Gaussian process classification. Journal of Machine Learning Research 9(Oct) 2035–2078 (2008)
17. Heskes, T., Zoeter, O.: Expectation propogation for approximate inference in dynamic bayesian networks. In: Uncertainty in Artificial Intelligence. 216–223 (2002)
18. Ypma, A., Heskes, T.: Novel approximations for inference in nonlinear dynamical systems using expectation propagation. Neurocomputing 69(1-3), 85–99 (2005)
19. Yu, B.M., Cunningham, J.P., Shenoy, K.V., Sahani, M.: Neural decoding of movements: From linear to nonlinear trajectory models. In: ICONIP (1). 586–595 (2007)
20. Vanhatalo, J., Pietiläinen, V., Vehtari, A.: Approximate inference for disease mapping with sparse Gaussian processes. Statistics in Medicine 29(15), 1580–1607 (2010)
21. Rue, H., Martino, S., Chopin, N.: Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society (Series B) 71(2), 319–392 (2009)
22. Cseke, B., Heskes, T.: Approximate marginals in latent Gaussian models. Journal of Machine Learning Research 12(Feb), 417–454 (2011)