Stochastic Collapsed Variational Inference for Structured Gaussian Process Regression Networks

Rui Meng, Herbert K. H. Lee, and Kristofer Bouchard

Abstract This paper presents an efficient variational inference framework for a family of structured Gaussian process regression network (SGPRN) models. We incorporate auxiliary inducing variables in latent functions and jointly treat both the distributions of the inducing variables and hyper-parameters as variational parameters. Then we take advantage of the collapsed representation of the model and propose structured variational distributions, which enables the decomposability of a tractable variational lower bound and leads to stochastic optimization. Our inference approach is able to model data in which outputs do not share a common input set, and with a computational complexity independent of the size of the inputs and outputs to easily handle datasets with missing values. Finally, we illustrate our approach on both synthetic and real data.

Keywords: stochastic optimization, Gaussian process, variational inference, multivariate time series, time-varying correlation

1 Introduction

Multi-output regression problems arise in various fields. Often, the processes that generate such datasets are nonstationary. Modern instrumentation has resulted in increasing numbers of observations, as well as the occurrence of missing values. This motivates the development of scalable methods for forecasting in such datasets.

Multi-ouput Gaussian process models or multivariate Gaussian process models (MGP) generalise the powerful Gaussian process predictive model to vector-valued

Rui Meng (✉) · Kristofer Bouchard
Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, USA,
e-mail: rmeng@lbl.gov;kebouchard@lbl.gov

Herbert K. H. Lee
University of California, Santa Cruz, USA, e-mail: herbie@ucsc.edu

P. Brito et al. (eds.), *Classification and Data Science in the Digital Age*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-031-09034-9_28

random fields [1]. Those models demonstrate improved prediction performance compared with independent univariate Gaussian processes (GP) because MGPs express correlations between outputs. Since the correlation information of data is encoded in the covariance function, modeling the flexible and computationally efficient cross-covariance function is of interest. In the literature of multivariate processes, many approaches are proposed to build valid cross-covariance functions including the linear model of coregionalization (LMC) [2], kernel convolution techniques [3], B-spline based coherence functions [4]. However, most of these models are designed for modelling low-dimensional stationary processes, and require Monte Carlo simulations, making inference in large datasets computationally intractable.

Modelling the complicated temporal dependencies across variables is addressed in [5, 6] by several adaptions of stochastic LMC. Such models can handle input-varying correlation across multivariate outputs. Especially for multivariate time series, [6] propose a SGPRN that captures time-varying scale, correlation, and smoothness. However, the inference in [6] is difficult to handle in applications where either the number of observations and dimension size are large or where missing data exist.

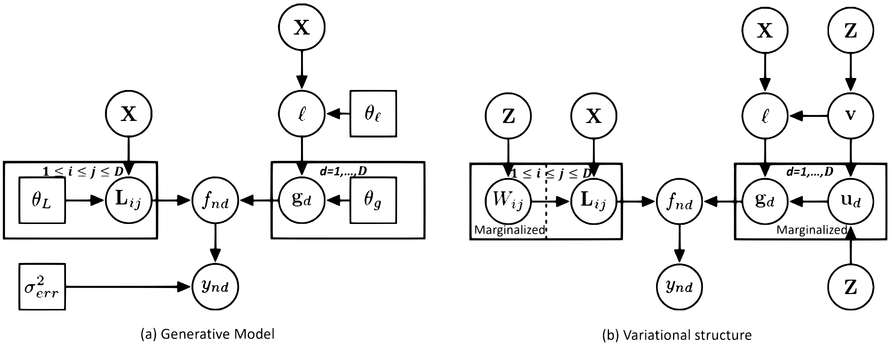
Here, we propose an efficient variational inference approach for the SGPRN by employing the inducing variable framework on all latent processes [7], taking advantage of its collapsed representation where nuisance parameters are marginalized out [8] and proposing a tractable variational bound amenable to doubly stochastic variational inference. We call our approach variational SGPRN (VSGPRN). This variational framework allows the model to handle missing data without increasing the computational complexity of inference. We numerically provide evidence of the benefits of simultaneously modeling time-varying correlation, scale and smoothness in both a synthetic experiment and a real-world problem.

The main contributions of this work are threefold:

- Learning structured Gaussian process regression networks using inducing variables on both mixing coefficients and latent functions.
- Employing doubly stochastic variational inference for structured Gaussian process regression networks by taking advantage of its collapsed representation and constructing a tractable lower bound of the loglikelihood, making it suitable for mini-batching learning.
- Demonstrating that our proposed algorithm succeeds in handling time-varying correlation on missing data under different scenarios in both synthetic data and real data.

2 Model

Assume $\mathbb{y}(\mathbb{x}) \in \mathbb{R}^D$ is a vector-valued function of $\mathbb{x} \in \mathbb{R}^P$, where D is the dimension size of the outputs and P is the dimension size of the inputs. SGPRN assumes that noisy observations $\mathbb{y}(\mathbb{x})$ are the linear combination of latent variables $\mathbb{g}(\mathbb{x}) \in \mathbb{R}^D$, corrupted by Gaussian noise $\epsilon(\mathbb{x})$. The coefficients $\mathbb{L}(\mathbb{x}) \in \mathbb{R}^{D \times D}$ of the latent functions are assumed to be a stochastic lower triangular matrix with



(a) Generative Model

(b) Variational structure

Fig. 1 Graphical model of VSGPRN. Left: Illustration of the generative model. Right: Illustration of the variational structure. The dashed (red) block means that we marginalize out those latent variables in the variational inference framework.

positive values on the diagonal for model identification [9, 6]. Thus, SGPRN is defined in the generative model of Figure 1 and it is $\mathbb{y}(\mathbb{x}) = \mathbb{f}(\mathbb{x}) + \epsilon(\mathbb{x})$, $\mathbb{f}(\mathbb{x}) = \mathbb{L}(\mathbb{x})\mathbb{g}(\mathbb{x})$ with independent white noise $\epsilon(\mathbb{x}) \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{err} I)$. We note that each latent function g_d in $\mathbb{g}$ is independently sampled from a GP with a non-stationary kernel K^g and the stochastic coefficients are modeled via a structured GP based prior as proposed in [9] with a stationary kernel K^l such that $g_d \overset{iid}{\sim} \mathrm{GP}(0, K^g)\,, d = 1, \dots, D\,,$ and $l_{ij} \sim \begin{cases} \mathrm{GP}(0, K^l)\,, & i > j\,, \\ \log\mathrm{GP}(0, K^l)\,, & i = j\,, \end{cases}$ where logGP denotes the log Gaussian process [10]. K^g is modelled as a Gibbs correlation function $K^g(\mathbb{x}, \mathbb{x}') = \sqrt{\frac{2\ell(\mathbb{x})\ell'(\mathbb{x})}{\ell(\mathbb{x})^2 + \ell(\mathbb{x}')^2}} \exp\left(-\frac{\|\mathbb{x} - \mathbb{x}'\|^2}{\ell(\mathbb{x})^2 + \ell(\mathbb{x}')^2}\right)\,, \ell \sim \log\mathrm{GP}(0, K^\ell)\,,$ where ℓ determines the input-dependent length scale of the shared correlations in K^g for all latent functions g_d. The varying length-scale process ℓ plays an important role in modelling nonstationary time series as illustrated in [11, 6].

Let $\mathbb{X} = \{\mathbb{x}_i\}_{i=1}^N$ be the set of observed inputs and $\mathbb{Y} = \{\mathbb{y}_i\}_{i=1}^N$ be the set of observed outputs. Denote η as the concatenation of all coefficients and all log length-scale parameters, i.e., $\eta = (\mathbb{l}, \tilde{\ell})$ evaluated at training inputs $\mathbb{X}$. Here, $\mathbb{l}$ is a vector including the entries below the main diagonal and the entries on the diagonal in the log scale and $\tilde{\ell} = \log \ell$ is the length-scale parameters in log scale. Also, denote $\theta = (\theta_l, \theta_\ell, \sigma^2_{err})$ as all hyper-parameters, where θ_l and θ_ℓ are the hyper-parameters in kernel K_l and K_ℓ. We note that directly inferring the posterior of the latent variables $p(\eta|\mathbb{Y}, \theta) \propto p(\mathbb{Y}|\eta, \sigma^2_{err}) p(\eta|\theta_l, \theta_\ell)$ is computationally intractable in general because the computational complexity of $p(\eta|\mathbb{Y}, \theta)$ is $O(N^3 D^3)$. To overcome this issue, we propose an efficient variational inference to significantly reduce the computational burden in the next section.

3 Inference

We introduce a shared set of inducing inputs $\mathbb{Z} = \{\mathbb{z}_m\}_{m=1}^{M}$ that lie in the same space as the inputs $\mathbb{X}$ and a set of shared inducing variables $\mathbb{w}_d$ for each latent function g_d evaluated at the inducing inputs $\mathbb{Z}$. Likewise, we consider inducing variables $\mathbb{u}_{ii}$ for the function $\log L_{ii}$ when $i = j$, $\mathbb{u}_{ij}$ for function L_{ij} when $i > j$, and inducing variables $\mathbb{v}$ for function $\log \ell(\mathbb{x})$ evaluated at inducing inputs $\mathbb{Z}$. We denote those collective variables as $\mathbb{l} = \{\mathbb{l}_{ij}\}_{i \geq j}$, $\mathbb{u} = \{\mathbb{u}_{ij}\}_{i \geq j}$, $\mathbb{g} = \{\mathbb{g}_d\}_{d=1}^{D}$, $\mathbb{w} = \{\mathbb{w}_d\}_{d=1}^{D}$, ℓ and $\mathbb{v}$. Then we redefine the model parameters $\eta = (\mathbb{l}, \mathbb{u}, \mathbb{g}, \mathbb{w}, \ell, \mathbb{v})$, and the prior of those model parameters is $p(\eta) = p(\mathbb{l}|\mathbb{w})p(\mathbb{w})p(\mathbb{g}|\mathbb{u}, \ell, \mathbb{v})p(\mathbb{u})p(\ell|\mathbb{v})p(\mathbb{v})$.

The core assumption of inducing point-based sparse inference is that the inducing variables are sufficient statistics for the training and testing data in the sense that the training and testing data are conditionally independent given the inducing variables. In the context of our model, this means that the posterior processes of L, g and ℓ are sufficiently determined by the posterior distribution of $\mathbb{u}$, $\mathbb{w}$ and $\mathbb{v}$. We propose a structured variational distribution and its corresponding variational lower bound. Due to the nonconjugacy of this model, instead of doing expectation in the evidence lower bound (ELBO), as is normally done in the literature, we perform the marginalization on inducing variables $\mathbb{u}$, $\mathbb{w}$ and $\mathbb{g}$, and then use the reparameterization trick to apply end-to-end training with stochastic gradient descent. We will also discuss a procedure for missing data inference and prediction.

To capture the posterior dependency between the latent functions, we propose a structured variational distribution of the model parameters η used to approximate its posterior distribution as $q(\eta) = p(\mathbb{l}|\mathbb{u})p(\mathbb{g}|\mathbb{w}, \ell, \mathbb{v})p(\ell|\mathbb{v})q(\mathbb{u}, \mathbb{w}, \mathbb{v})$. This variational structure is illustrated in Figure 1. The variational distribution of the inducing variables $q(\mathbb{u}, \mathbb{w}, \mathbb{v})$ fully characterizes the distribution of $\mathbb{q}(\eta)$. Thus, the inference of $q(\mathbb{u}, \mathbb{w}, \mathbb{v})$ is of interest. We assume the parameters $\mathbb{u}$, $\mathbb{w}$, and $\mathbb{v}$ are Gaussian and mutually independent.

Given the definition of Gaussian process priors for the SGPRN, the conditional distributions $p(\mathbb{l}|\mathbb{u})$, $p(\mathbb{g}|\mathbb{w}, \tilde{\ell}, \mathbb{v})$, and $p(\ell|\mathbb{v})$ have closed-form expressions and all are Gaussian, except for $p(\ell|\mathbb{v})$, which is log Gaussian. The ELBO of the log likelihood of observations under our structured variational distribution $q(\eta)$ is derived using Jensen's inequality as:

$$\log p(\mathbb{Y}) \geq E_{q(\eta)}\left[\log\left(\frac{p(\mathbb{Y}|\mathbb{g}, \mathbb{l})p(\mathbb{u})p(\mathbb{w})p(\mathbb{v})}{q(\mathbb{u}, \mathbb{w}, \mathbb{v})}\right)\right] = R + A\,, \tag{1}$$

where $R = \sum_{n=1}^{N}\sum_{d=1}^{D} E_{q(\mathbb{g}_n, \mathbb{l}_n)} \log(p(y_{nd}|\mathbb{g}_n, \mathbb{l}_n))$ is the reconstruction term and $A = \mathrm{KL}(q(\mathbb{u})||p(\mathbb{u})) + \mathrm{KL}(q(\mathbb{w})||p(\mathbb{w})) + \mathrm{KL}(q(\mathbb{v})||p(\mathbb{v}))$ is the regularization term. $\mathbb{g}_n = \{g_{dn} = (\mathbb{g}_d)_n\}_{d=1}^{D}$ and $\mathbb{l}_n = \{l_{ijn} = (\mathbb{l}_{ij})_n\}_{i \geq j}$ are latent variables.

The structured decomposition trick for $q(\eta)$ has also been used by [12] to derive variational inference for the multivariate output case. The benefit of this structure is that all conditional distributions in $q(\eta)$ can be cancelled in the derivation of the lower bound in (1), which alleviates the computational burden of inference. Because of the conditional independence of the reconstruction term in (1) given $\mathbb{g}$ and $\mathbb{l}$, the lower bound decomposes across both inputs and outputs and this enables the use of stochastic optimization methods. Moreover, due to the Gaussian assumption in the prior and variational distributions of the inducing variables, all KL divergence terms in the regularization term A are analytically tractable. Next, instead of directly computing expectation, we leverage stochastic inference [13].

Stochastic inference requires sampling of $\mathbb{l}$ and $\mathbb{g}$ from the joint variational posterior $q(\eta)$. Directly sampling them would introduce much uncertainty from intermediate variables and thus make inference inefficient. To tackle this issue, we marginalize unnecessary intermediate variables $\mathbb{u}$ and $\mathbb{w}$ and obtain the marginal distributions $q(\mathbb{l}) = \prod_{i=j} \log \mathcal{N}(\mathbb{l}_{ii}|\tilde{\mu}^l_{ii}, \tilde{\Sigma}^l_{ii}) \prod_{i>j} \mathcal{N}(\mathbb{l}_{ij}|\tilde{\mu}^l_{ij}, \tilde{\Sigma}^l_{ij})$ and $q(\mathbb{g}|\ell, \mathbb{v}) = \prod_{d=1}^{D} \mathcal{N}(\mathbb{g}_d|\tilde{\mu}^g_d, \tilde{\Sigma}^g_d)$ with a joint distribution $q(\ell, \mathbb{v}) = p(\ell|\mathbb{v})q(\mathbb{v})$, where the conditional mean and covariance matrix are easily derived. The corresponding marginal distributions $q(\mathbb{l}_n)$ and $q(\mathbb{g}_n|\ell, \mathbb{v})$ at each n are also easy to derive. Moreover, we conduct collapsed inference by marginalizing the latent variables $\mathbb{g}_n$, so then the individual expectation is

$$\mathrm{E}_{q(\mathbb{g}_n, \mathbb{l}_n)} \log(p(y_{nd}|\mathbb{g}_n, \mathbb{l}_n)) = \int (L_{nd}) q(\ell_n, \mathbb{v}) q(\mathbb{l}_{d \cdot n}) d(\mathbb{l}_{d \cdot n}, \ell_n, \mathbb{v})), \quad (2)$$

where $L_{nd} = \log \mathcal{N}(y_{nd}| \sum_{j=1}^{D} l_{djn} \tilde{\mu}^g_{jn}, \sigma^2_{err}) - \frac{1}{2\sigma^2_{err}} \sum_{j=1}^{D} l^2_{djn} \tilde{\sigma}^{g2}_{jn}$ measure the reconstruction performance for observations $\mathbb{y}_{nd}$.

Directly evaluating the ELBO is still challenging due to the non-linearities introduced by our structured prior. Recent progress in black box variational inference [13] avoids this difficulty by computing noisy unbiased estimates of the gradient of ELBO, via approximating the expectations with unbiased Monte Carlo estimates and relying on either score function estimators [14] or reparameterization gradients [13] to differentiate through a sampling process. Here we leverage the reparameterization gradients for stochastic optimization for model parameters. We note that evaluating ELBO (1) involves two sources of stochasticity from Monte Carlo sampling in (2) and from data sub-sampling stochasticity [15]. The prediction procedure is based on Bayes' rule and replaces the posterior distribution by the inferred variational distribution. In the case of missing data, the only modification in (1) is in the reconstruction term, where we sum up the likelihoods of observed data instead of complete data.

4 Experiments

This section illustrates the performance of our model on multivariate time series. We first show that our approach can model the time-varying correlation and smoothness of outputs on 2D synthetic datasets in three scenarios with respect to different types of frequencies but the same missing data mechanism. Then, we compare the imputation performance on missing data with other inducing-variable based sparse multivariate Gaussian process models on a real dataset.

We conduct experiments on three synthetic time series with low frequency (LF), high frequency (HF) and varying frequency (VF) respectively. They are generated from the system of equations $y_1(t) = 5\cos(2\pi w t^s) + \epsilon_1(t), y_2(t) = 5(1-t)\cos(2\pi w t^s) - 5t\cos(2\pi w t^s) + \epsilon_2(t)$, where $\{\epsilon_i(t)\}_{i=1}^2$ are independent standard white noise processes. The value of w refers to the frequency and the value of s characterizes the smoothness. The LF and HF datasets use the same $s = 1$, implying the smoothness is invariant across time. But they employ different frequencies, $w = 2$ for LF and $w = 5$ for HF (i.e., two periods and five periods in a unit time interval respectively). The VF dataset takes $s = 2$ and $w = 5$, so that the frequency of the function is gradually increasing as time increases. For all three datasets, the system shows that as time t increases from 0 to 1, the correlation between $y_1(t)$ and $y_2(t)$ gradually varies from positive to negative. Within each dataset, we randomly select 200 training data points, in which 100 time stamps are sampled on the interval $(0, 0.8)$ for the first dimension and the other 100 time stamps sampled on the interval $(0.2, 1)$ for the second dimension. For the test inputs, we randomly select 100 time stamps on the interval $(0, 1)$ for each dimension.

Table 1 Prediction measurements on three synthetic datasets and different models. LF, HF and VF refer to low-frequency, high-frequency, and time-varying datasets. Three prediction measures are root mean square error (RMSE), average length of confidence interval (ALCI), and coverage rate (CR). All three measurements are summarized by the mean and standard deviation across 10 runs with different random initializations.

Data	Model	RMSE	ALCI	CR
LF	IGPR [16]	2.25(1.33e-13)	2.18(1.88e-13)	0.835(0)
	ICM [17]	2.26(2.54e-5)	2.18(1.22e-5)	0.835(0)
	CMOGP [12]	1.43(6.12e-2)	1.36(1.98e-1)	0.651(3.00e-2)
	VGPRN [18]	1.01(0.31)	-	-
	VSGPRN	**1.00(1.43e-1)**	2.21(6.56e-2)	**0.892(1.63e-2)**
HF	IGPR [16]	1.51(6.01e-14)	3.17(1.30e-13)	0.915(2.22e-16)
	ICM [17]	1.52(1.01e-5)	3.17(1.19e-5)	0.910(0)
	CMOGP [12]	1.29(3.04e-2)	2.34(3.31e-1)	0.729(3.07e-2)
	VGPRN [18]	1.11(0.25)	-	-
	VSGPRN	**1.10(1.98e-1)**	2.74(7.94e-2)	**0.930(1.14e-2)**
VF	IGPR [16]	1.64(8.17e-14)	3.19(3.02e-13)	0.875(0)
	ICM [17]	1.66(2.37e-3)	3.16(1.49e-3)	0.880(1.50e-3)
	CMOGP [12]	2.24(3.08e-1)	2.56(9.29e-1)	0.697(1.56e-1)
	VGPRN [18]	1.04(0.67)	-	-
	VSGPRN	**1.24(1.33e-1)**	2.92(1.21e-1)	**0.887(9.80e-3)**

We quantify the model performance in terms of root mean square error (RMSE), average length of confidence interval (ALCI), and coverage rate (CR) on the test set. A smaller RMSE corresponds to better predictive performance of the model, and a smaller ALCI implies a smaller predictive uncertainty. As for CR, the better the model prediction performance is, the closer CR is to the percentile of the credible band. Those results are reported by the mean and standard deviation with 10 different random initializations of model parameters. Quantitative comparisons relating to all three datasets are in Table 1. We compare with independent Gaussian process regression (IGPR) [16], the intrinsic coregionalization model (ICM) [17], Collaborative Multi-Output Gaussian Processes (CMOGP) [12] and variational inference of Gaussian process regression networks [18] on three synthetic datasets. In both CMOGP and VSGPRN approaches, we use 20 inducing variables. We further examined model predictive performance on a real-world dataset, the PM2.5 dataset from the UCI Machine Learning Repository [19]. This dataset tracks the concentration of fine inhalable particles hourly in five cities in China, along with meteorological data, from Jan 1st, 2010 to Dec 31st, 2015. We compare our model with two sparse Gaussian process models, i.e., independent sparse Gaussian process regression (ISGPR) [20] and the sparse linear model of corregionalization (SLMC) [17]. In the dataset, we consider six important attributes and use 20% of the first 5000 standardized multivaritate for training and use the others for testing. The RMSEs on the testing data are shown in Table 2, illustrating that VSGPRN had better prediction performance compared with ISGPR and SLMC, even when using fewer inducing points.

Table 2 Empirical results for PM2.5 dataset. Each model's performance is summarized by its RMSE on the testing data. The number of equi-spaced inducing points is given in parentheses.

Data	ISGPR (100) [20]	SLMC (100) [17]	VSGPRN (50)	VSGPRN (100)	VSGPRN (200)
PM2.5	0.994	0.948	0.840	0.708	0.625

5 Conclusions

We propose a novel variational inference approach for structured Gaussian process regression networks named the variational structured Gaussian process regression network, VSGPRN. We introduce inducing variables and propose a structured variational distribution to reduce the computational burden. Moreover, we take advantage of the collapsed representation of our model and construct a tractable lower bound of the log likelihood to make it suitable for doubly stochastic inference and easy to handle missing data. In our method, the computation complexity is independent of the size of the inputs and the outputs. We illustrate the superior predictive performance for both synthetic and real data.

Our inference approach, VSGPRN can be widely used for high dimensional time series to model complicated time-varying dependence across multivariate outputs. Moreover, due to its scalability and flexibility, it can be widely applied for irregularly sampled incomplete large datatsets that widely exist in various research fields including healthcare, environmental science and geoscience.

References

1. Álvarez, M., Lawrence, N.: Computationally efficient convolved multiple output Gaussian processes. J. Mach. Learn. Res. **12**, 1459-1500 (2011)
2. Goulard, M., Voltz, M.: Linear coregionalization model: tools for estimation and choice of cross-variogram matrix. Math. Geol. **24**, 269-286 (1992)
3. Gneiting, T., Kleiber, W., Schlather, M.: Matérn cross-covariance functions for multivariate random fields. J. Am. Stat. Assoc. **105**, 1167-1177 (2010)
4. Qadir, G., Sun, Y.: Semiparametric estimation of cross-covariance functions for multivariate random fields. Biom. **77**, 547-560 (2021)
5. Gelfand, A., Schmidt, A., Banerjee, S., Sirmans, C.: Nonstationary multivariate process modeling through spatially varying coregionalization. Test. **13**, 263-312 (2004)
6. Meng, R., Soper, B., Lee, H., Liu, V., Greene, J., Ray, P.: Nonstationary multivariate Gaussian processes for electronic health records. J. Biom. Inform. **117**, 103698 (2021)
7. Titsias, M., Lawrence, N.: Bayesian Gaussian process latent variable model. Int. Conf. Artif. Intell. Stat. 844-851 (2010)
8. Teh, Y., Newman, D., Max Welling, M.: A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In: Schölkopf, B., Platt, J., Hofmann, T. (eds.) Advances in Neural Information Processing Systems **19**, (2006)
9. Guhaniyogi, R., Finley, A., Banerjee, S., Kobe, R.: Modeling complex spatial dependencies: Low-rank spatially varying cross-covariances with application to soil nutrient data. J. Agric. Biol. Environ. Stat. **18**, 274-298 (2013)
10. Møller, J., Syversveen, A., Waagepetersen, R.: Log Gaussian Cox processes. Scand. J. Stat. **25**, 451-482 (1998)
11. Remes, S., Heinonen, M., Kaski, S.: Non-stationary spectral kernels. Adv. Neural Inf. Process. Syst. **30** (2017), https://proceedings.neurips.cc/paper/2017/file/c65d7bd70fe3e5e3a2f3de681edc193d-Paper.pdf
12. Nguyen, T., Bonilla, E., et al.: Collaborative multi-output Gaussian processes. Uncertain. Artif. Intell. 643-652 (2014)
13. Titsias, M., Lázaro-Gredilla, M.: Doubly stochastic variational Bayes for non-conjugate inference. Int. Conf. Mach. Learn. 1971-1979 (2014)
14. Ranganath, R., Gerrish, S., Blei, D.: Black box variational inference. Int. Conf. Artif. Intell. Stat. 814-822 (2014)
15. Hoffman, M., Blei, D., Wang, C., Paisley, J.: Stochastic variational inference. J. Mach. Learn. Res. **14**, 1303-1347 (2013)
16. Rasmussen, C., Kuss, M.: Gaussian processes in reinforcement learning. Adv. Neural Inf. Process. Syst. 751-759 (2004)
17. Wackernagel, H.: Multivariate geostatistics: an introduction with applications. Springer Science & Business Media (2013)
18. Nguyen, T., Bonilla, E.: Efficient variational inference for Gaussian process regression networks. Int. Conf. Artif. Intell. Stat. 472-480 (2013)
19. Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., Huang, H., Chen, S.: Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC and winter heating. Proc. R. Soc. A: Math. Phys. Eng. Sci. **471**, 20150257 (2015) https://royalsocietypublishing.org/doi/abs/10.1098/rspa.2015.0257
20. Snelson, E., Ghahramani, Z.: Sparse Gaussian processes using pseudo-inputs. Adv. Neural Inf. Process. Syst. 1257-1264 (2006), http://papers.nips.cc/paper/2857-sparse-gaussian-processes-using-pseudo-inputs.pdf