Design of Covariance Functions using Inter-Domain Inducing Variables

Felipe Tobar
ftobar@dim.uchile.cl
Center for Mathematical Modeling
Universidad de Chile

Thang D. Bui
tdb40@cam.ac.uk
Department of Engineering
University of Cambridge

Richard E. Turner
ret26@cam.ac.uk
Department of Engineering
University of Cambridge

Abstract

We introduced the Gaussian Process Convolution Model (GPCM) in [1], a time-series model for stationary signals based on the convolution between a continuous-time white-noise process and a continuous-time linear filter drawn from Gaussian process. The GPCM is, conditionally, itself a Gaussian process with a nonparametric kernel defined in a probabilistic fashion. Learning is achieved using a variational free-energy approach based on inter-domain inducing variables that summarise the (posterior) continuous-time linear filter and the driving white-noise process. However, the inter-domain transformation in [1] considers local averages of the noise process and therefore requires a large number inducing variables to represent underlying functions of complex spectra. In this paper, we develop an alternative transformation operating directly in the frequency domain, that retains the same modelling and predictive power as the original but requires fewer inducing variables and, consequently, has a reduced training time. The proposed approach is validated in a spectrum estimation task on a real-world time series.

Note: this is a short version of [1] with additional results. For more technical details and experiments please refer to the original paper.

1 Introduction

Gaussian process (GP) regression models have become a standard tool in Bayesian signal estimation due to their expressiveness, robustness to overfitting and tractability [2]. GPs are general priors over unknown nonlinear functions, which are updated using the observed data to produce a posterior over the functions of interest. The form of the covariance function (or kernel) of the GP is arguably the central modelling choice, as it encapsulates *a priori* assumptions about the unknown function, such as smoothness, stationarity or periodicity. Recently, sophisticated automated approaches to kernel design have been developed that construct kernel mixtures on the basis of incorporating different measures of similarity [3, 4], or more generally by both adding and multiplying kernels, thus mimicking the way in which a human would search for the best kernel [5]. Alternatively, a flexible parametric kernel can be used as in the case of the spectral mixture kernels, where the power spectral density (PSD) of the GP is parametrised by a mixture of Gaussians [6]. However, the complexity of the kernels that can be designed in this way is hindered by the computational intractability when searching in the large kernel space and potentially the overfitting problem when dealing with a large number of hyperparameters.

We are interested in designing expressive and tractable covariance functions for real-world time series. In [1], we introduced the Gaussian Process Convolution Model, which expresses time series as the output of a linear and time-invariant system defined by a convolution between a white-noise process and a continuous-time linear filter. By considering the filter to be drawn from a GP, the expected second-order statistics (and, as a consequence, the spectral density) of the output signal are defined in a nonparametric fashion. That is, the kernels themselves are treated nonparametrically

to enable flexible forms whose complexity can grow as more structure is revealed in the data. We also developed a tractable variational free-energy approach based on inter-domain inducing points in [1]; however, such approach needs a large number of inducing points to produce accurate representations, resulting in a large computational complexity and training times. This paper investigates an alternative inter-domain transformation that operates directly in the frequency domain and can retain the same predictive performance using a much smaller number of inducing points.

2 Regression model: Convolving a linear filter and a white-noise process

The Gaussian Process Convolution Model (GPCM) [1] can be viewed as constructing a distribution over functions $f(t)$ using a two-stage generative model. In the first stage, a continuous filter function $h(t) : \mathbb{R} \mapsto \mathbb{R}$ is drawn from a GP with covariance function $\mathcal{K}_h(t_1, t_2)$. In the second stage, the function $f(t)$ is produced by convolving the filter with continuous time white-noise $x(t)$. The white-noise can be treated informally as a draw from a GP with a delta-function covariance,

$$h(t) \sim \mathcal{G}P(\mathbf{0}, \mathcal{K}_h(t_1, t_2)), \quad x(t) \sim \mathcal{G}P(\mathbf{0}, \sigma_x^2 \delta(t_1 - t_2)), \quad f(t) = \int_{\mathbb{R}} h(t-\tau)x(\tau)\mathrm{d}\tau. \tag{1}$$

Real-world signals have finite power (thus ensuring stability of the system) and potentially complex spectral content. To fulfil these conditions, we model the linear filter $h(t)$ as a draw from a squared exponential GP that is multiplied by a Gaussian window (centred on zero) in order to restrict its extent. The resulting *decaying squared exponential* (DSE) covariance function is given by a squared exponential (SE) covariance pre- and post-multiplied by $e^{-\alpha t_1^2}$ and $e^{-\alpha t_2^2}$ respectively, that is,

$$\mathcal{K}_h(t_1, t_2) = K_{\text{DSE}}(t_1, t_2) = \sigma_h^2 e^{-\alpha t_1^2} e^{-\gamma(t_1 - t_2)^2} e^{-\alpha t_2^2}, \ \alpha, \gamma, \sigma_h > 0. \tag{2}$$

The DSE model for the filter $h(t)$ provides a flexible prior distribution over linear systems, where the hyperparameters have physical meaning: σ_h^2 controls the power of the output $f(t)$; $1/\sqrt{\gamma}$ is the characteristic timescale of the filter that, in turn, determines the typical frequency content of the system; finally, $1/\sqrt{\alpha}$ is the temporal extent of the filter which controls the length of time correlations in the output signal and, equivalently, the bandwidth characteristics in the frequency domain. With this choice of prior for the filter h, we refer to the proposed model as DSE-GPCM.

3 Inference and learning using variational methods

A variational free-energy approach is proposed to learn the filter $h(t)$ (system identification in the control community [7]) and infer the white-noise process $x(t)$ from a noisy dataset $\mathbf{y} \in \mathbb{R}^N$ produced by their convolution and additive Gaussian noise, $y(t) = f(t) + \epsilon(t) = \int_{\mathbb{R}} h(t-\tau)x(\tau)\mathrm{d}\tau + \epsilon(t)$, $\epsilon(t) \sim \mathcal{N}(0, \sigma_\epsilon^2)$. A key ingredient of the proposed approach is the use of a finite set of inter-domain inducing variables which can summarise the underlying latent functions. In order to form the variational inter-domain approximation, we first expand the model with additional variables. We use X to denote the set of all integral transformations of $x(t)$ with members $u_x(t) = \int w(t,\tau)x(\tau)\mathrm{d}\tau$ (which includes the original white-noise process when $w(t,\tau) = \delta(t-\tau)$) and identically define the set H with members $u_h(t) = \int w(t,\tau)h(\tau)\mathrm{d}\tau$. We then choose a structured mean-field variational distribution $q(H, X)$ that mirrors the form of the joint distribution,

$$\begin{aligned} p(\mathbf{y}, H, X) &= p(x|\boldsymbol{u_x})p(h|\boldsymbol{u_h})p(\boldsymbol{u}_x)p(\boldsymbol{u}_h)p(\mathbf{y}|h,x) \\ q(H, X) &= p(x|\boldsymbol{u}_x)p(h|\boldsymbol{u}_h)q(\boldsymbol{u}_x)q(\boldsymbol{u}_h) = q(H)q(X). \end{aligned}$$

Critically this leads to a tractable variational lower bound of the model evidence:

$$\mathcal{F} = \int q(h, x, \boldsymbol{u}_h, \boldsymbol{u}_x) \log \frac{p(\mathbf{y}|h,x)p(\boldsymbol{u}_h)p(\boldsymbol{u}_x)}{q(\boldsymbol{u}_h)q(\boldsymbol{u}_x)} \mathrm{d}h\mathrm{d}x\mathrm{d}\boldsymbol{u}_h\mathrm{d}\boldsymbol{u}_x \tag{3}$$

$$= \mathbb{E}_q \left[\log p(\mathbf{y}|h,x)\right] - \mathrm{KL}[q(\boldsymbol{u}_h)||p(\boldsymbol{u}_h)] - \mathrm{KL}[q(\boldsymbol{u}_x)||p(\boldsymbol{u}_x)]. \tag{4}$$

3.1 Choice of the inducing variables u_h and u_x

In order to choose the domain of the inducing variables, it is useful to consider inference for the white-noise process given a fixed window $h(t)$. Typically, we assume that the window $h(t)$ is

smoothly varying and can be reconstructed to a desired degree of accuracy by taking a finite set of observations. As such, we choose the inducing variables u_h directly in the domain of the $filter$. However, this is not feasible for u_x since a continuous-time white-noise process contains power at all frequencies and hence cannot be reconstructed from a finite number of observations.

In [1] we opted for modelling the inducing points u_x as local (spatial) averages of the white noise given by $u_x = \sigma \int_{\mathbb{R}} \exp\left(-\frac{1}{2l^2}(t_x - \tau)^2\right) x(\tau)\mathrm{d}\tau$, that is, a low-pass version of the white noise, where the frequency is controlled by the distance between the locations t_x's. The rationale behind this choice is the low-pass property of the filter, which means that only low frequency content of the white noise is present in the output signal. A drawback of this formulation is that, when the filter does allow higher frequencies of white noise to go through, the distance between t_x's has to be short and therefore a large number of inducing points u_x are required to retain an accurate representation of the signal.

An alternative approach to form the inter-domain inducing points that can recover higher (or in fact, targeted) frequencies without the need of placing too many points t_x's is to use a complex-exponential integral transformation of the following form

$$u_x = \left[\sigma \int_{\mathbb{R}} e^{-\frac{1}{2l^2}(t_x-\tau)^2} \cos(\omega_x \tau) x(\tau)\mathrm{d}\tau,\ \sigma \int_{\mathbb{R}} e^{-\frac{1}{2l^2}(t_x-\tau)^2} \sin(\omega_x \tau) x(\tau)\mathrm{d}\tau\right]. \tag{5}$$

These variables, referred to as *harmonic inducing points*, correspond to a band-pass approximation of the original white noise process around the inducing locations t_x's and therefore allow for extracting information of the underlying process around the frequencies ω_x. We did not optimise the bandwidths ω_x and they were chosen using heuristics.

4 Experiments

The DSE-GPCM was tested using synthetic data with known statistical properties and real-world signals. The experiment set-up was similar to [1], please refer to the paper for more details.

4.1 Learning known parametric kernels

We first generated a synthetic dataset from a GP with a spectral mixture kernel (GP-SM) and unit-variance Gaussian observation noise. Fig. 1 shows that the proposed DSE-GPCM obtained remarkably accurate estimates of the underlying SM kernel and its spectral density, and as a result, performed within a 5% of the true model in a missing data imputation task.

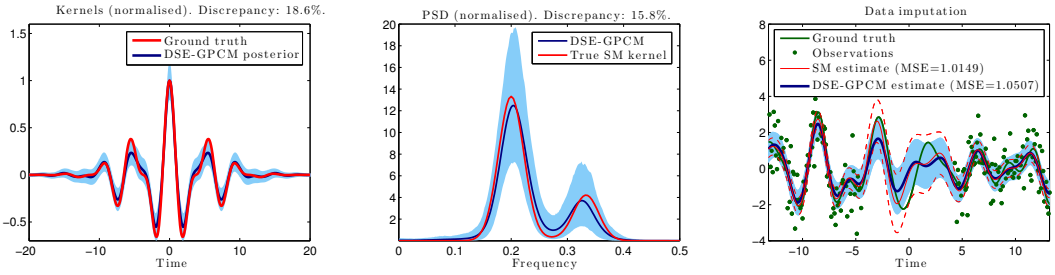

Figure 1: Joint learning of an SM kernel and data imputation using a nonparametric kernel. True and learnt kernel (left), true and learnt spectra (centre) and data imputation region (right).

4.2 Learning the spectrum of real-world audio signal

We considered a 1750-sample audio signal from the TIMIT corpus. We compared the proposed model and inference method against (i) the spectral mixture kernel (GP-SM) [6], and (ii) tracking the Fourier coefficients using a Kalman filter (Kalman-Fourier [8]), using only 20% of the data (unevenly-sampled). Additionally, both the Yule-Walker method and the periodogram [9] were considered as benchmarks using all the available data. We experimented with the spatial (low pass) from [1] and the harmonic (band pass) inter-domain transformations for the DSE-GPCM described

in sec. 3.1. The number of inducing points for the filter and white-noise were $N_{\mathbf{h}} = 100\, N_{\mathbf{x}} = 151$ for the spatial transformation, and $N_{\mathbf{h}} = 32,\ N_{\mathbf{x}} = 51$ for the harmonic case.

The estimates of the PSD (top) and the learnt kernel (bottom) using the proposed DSE-GPCM and other methods are shown in fig. 2. Notice that despite the fewer inducing points, the harmonic-inter-domain transformation provideed better estimates (w.r.t. to fitting the periodogram and the autocorrelation function) as compared to the spatial approach. Also, by using about one third of the inducing variables, the harmonic approach resulted in a training time of 180 seconds, whereas the spatial one required 920 seconds — both until convergence using the BFGS algorithm.

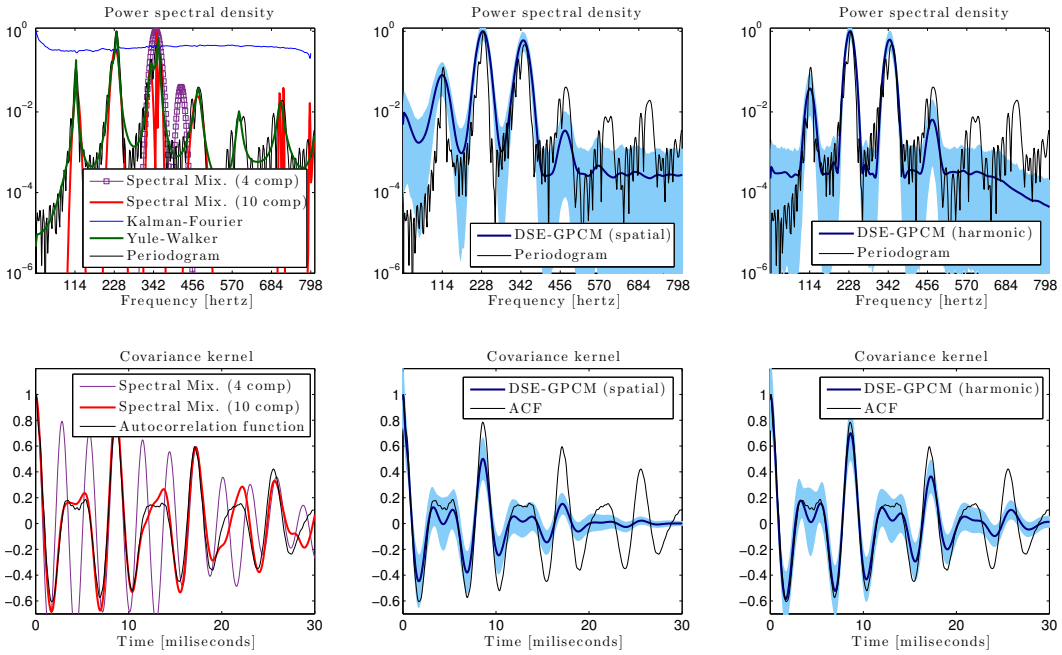

Figure 2: Comparison of GP-SM, Kalman-Fourier, Yule-Walker, raw periodogram (left), and our proposed DSE-GPCM approach using spatial (centre) and harmonic (right) inter-domain inducing variables. The PSD estimates are shown at the top and the kernel estimates at the bottom.

5 Discussion

The Gaussian Process Convolution Model (GPCM) expresses stationary time series by a convolution between a filter function and a white-noise process. Learning the model from data is achieved by finding the approximate posterior of finite representations of the filter and noise process using variational free-energy methods; this allows for predicting the unknown signal and inferring both the covariance kernel and the spectrum in a probabilistic, analytically and computationally tractable manner. We have extended the (spatial) inter-domain transformation proposed in [1] with a harmonic one that maintains modelling and predictive performance using fewer inducing variables, thus resulting in a reduction of about 80% of the training time. The proposed approach was validated in the recovery of spectral density from non-uniformly sampled time series of a real-world audio signal.

Future research directions include: (i) automated discovery of the frequency of the harmonic transformation (ii), scaling the proposed approach to work on longer time series, for instance, through the use of tree-structured approximations [10], (iii) operation on higher-dimensional input spaces, e.g., by means of a factorisation of the latent kernel, whereby the number of inducing points for the filter only increases linearly with the dimension, rather than exponentially.

Acknowledgements

Part of this work was carried out when F.T. was with the University of Cambridge. F.T. thanks CONICYT-PAI grant 82140061 and Basal-CONICYT Center for Mathematical Modeling (CMM). R.T. thanks EPSRC grants EP/L000776/1 and EP/M026957/1. T.B. thanks Google.

References

[1] F. Tobar, T. D. Bui, and R. E. Turner, "Learning stationary time series using Gaussian processes with nonparametric kernels," in *Advances in Neural Information Processing Systems 29*, 2015.

[2] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[3] M. Gönen and E. Alpaydin, "Multiple kernel learning algorithms," *The Journal of Machine Learning Research*, vol. 12, pp. 2211–2268, 2011.

[4] F. Tobar, S.-Y. Kung, and D. Mandic, "Multikernel least mean square algorithm," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 25, no. 2, pp. 265–277, 2014.

[5] D. Duvenaud, J. R. Lloyd, R. Grosse, J. B. Tenenbaum, and Z. Ghahramani, "Structure discovery in nonparametric regression through compositional kernel search," in *Proc. of International Conference on Machine Learning*, pp. 1166–1174, June 2013.

[6] A. G. Wilson and R. P. Adams, "Gaussian process kernels for pattern discovery and extrapolation," in *Proc. of International Conference on Machine Learning*, 2013.

[7] A. H. Jazwinski, *Stochastic processes and filtering theory*. New York, Academic Press., 1970.

[8] Y. Qi, T. Minka, and R. W. Picara, "Bayesian spectrum estimation of unevenly sampled nonstationary data," in *Proc. of IEEE ICASSP*, vol. 2, pp. II–1473–II–1476, 2002.

[9] D. B. Percival and A. T. Walden, *Spectral Analysis for Physical Applications*. Cambridge University Press, 1993. Cambridge Books Online.

[10] T. D. Bui and R. E. Turner, "Tree-structured Gaussian process approximations," in *Advances in Neural Information Processing Systems 27*, pp. 2213–2221, 2014.