arXiv:2203.06997v1 [stat.ML] 14 Mar 2022

Modelling Non-Smooth Signals with Complex Spectral Structure

Wessel P. Bruinsma
University of Cambridge
Invenia Labs
wpb23@cam.ac.uk

Martin Tegnér*
University of Oxford
Oxford-Man Institute
mt@robots.ox.ac.uk

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

Abstract

The Gaussian Process Convolution Model (GPCM; Tobar et al., 2015a) is a model for signals with complex spectral structure. A significant limitation of the GPCM is that it assumes a rapidly decaying spectrum: it can only model smooth signals. Moreover, inference in the GPCM currently requires (1) a mean-field assumption, resulting in poorly calibrated uncertainties, and (2) a tedious variational optimisation of large covariance matrices. We redesign the GPCM model to induce a richer distribution over the spectrum with relaxed assumptions about smoothness: the Causal Gaussian Process Convolution Model (CGPCM) introduces a causality assumption into the GPCM, and the Rough Gaussian Process Convolution Model (RGPCM) can be interpreted as a Bayesian nonparametric generalisation of the fractional Ornstein–Uhlenbeck process. We also propose a more effective variational inference scheme, going beyond the mean-field assumption: we design a Gibbs sampler which directly samples from the optimal variational solution, circumventing any variational optimisation entirely. The proposed variations of the GPCM are validated in experiments on synthetic and real-world data, showing promising results.

1 INTRODUCTION

Gaussian processes (GPs) form a popular and powerful probabilistic framework for modelling functions (Rasmussen and Williams, 2006). They are successfully applied in a wide variety of contexts and are state of the art in numerous regression tasks (Bui et al., 2016). Gaussian processes are nonparametric models that grow in complexity as more data is observed, which makes them robust against overfitting. They achieve this automatic calibration of complexity by posing a prior distribution directly over the underlying function $f \colon \mathcal{T} \to \mathbb{R}$. In particular, the defining property of a Gaussian process is that any finite collection of function values $f(t_1), \ldots, f(t_n)$ is multivariate Gaussian distributed.

The key modelling decision when using Gaussian processes is the choice of covariance function $k(t, t') = \mathrm{cov}(f(t), f(t'))$, also called the *kernel*. The kernel encodes prior information about the underlying function f. For example, the kernel specifies the smoothness of f and the typical length scale on which f varies. A kernel is *stationary* if it only depends on the difference of its arguments: $k(t, t') = k(t - t')$. In that case, if the data is translated, the predictions are translated accordingly, a symmetry called *translation equivariance* which is often desirable. A stationary kernel is characterised by its Fourier transform—a fact known as Bochner's theorem—where the Fourier transform is called the *power spectral density* (PSD) or simply spectrum. If f is decomposed into complex exponentials with random amplitudes, then the spectrum tells us how the variances of these random amplitudes vary with frequency.

A popular choice for the kernel is the *exponentiated quadratic* (EQ) kernel: $k(t, t') = \exp\left(-\frac{1}{2\ell^2}\|t - t'\|^2\right)$. The EQ kernel assumes that f is infinitely differentiable and varies on only a single length scale ℓ. Whilst appropriate for many tasks, these assumptions are too rigid for harder regression problems, which require more expressive kernels. From the perspective of the spectrum, the EQ kernel assumes that the spectrum of f is necessarily of the form $\mathrm{PSD}(\omega) = c_1 e^{-c_2 \omega^2}$. This is restrictive, because real-world signals often have much richer spectral structure.

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s). *Currently with the University of Copenhagen and Oxford-Man Institute, University of Oxford.

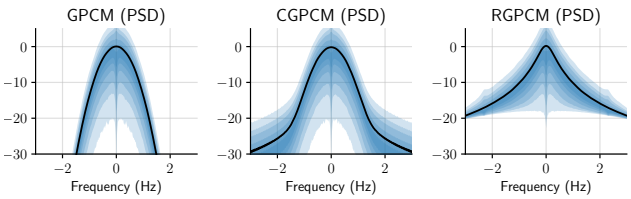


Figure 1: Visualisation of the nonparametric prior over the power spectral density by the GPCM (left) and the two variants of the GPCM introduced in this paper (middle and right). Observe that the GPCM has a quickly decaying spectrum, whereas the CGPCM and especially the RGPCM have support at higher frequencies.

An important property of the spectrum is the behaviour at high frequencies. Specifically, the asymptotic decay of $\mathrm{PSD}(\omega)$ is intimately connected to the regularity of sample paths of f. For example, sample paths of f are n-times differentiable if and only if the $2n^{\text{th}}$ spectral moment is finite, $\int \omega^{2n} \mathrm{PSD}(\omega)\, \mathrm{d}\omega < \infty$ (Theorem 4, Cambanis, 1973). Since the PSD of the EQ kernel $\mathrm{PSD}(\omega) = c_1 e^{-c_2 \omega^2}$ decays quicker than any polynomial, sample paths of a Gaussian process with an EQ kernel are infinitely differentiable.

Developing flexible and expressive kernels as well as choosing the right kernel for a particular task is an active area of research. Many approaches let the kernel be of a flexible parametric form (Wilson and Adams, 2013; Calandra et al., 2016; Sun et al., 2018) or search over a large space of kernels (Grosse et al., 2012; Duvenaud, 2014; Malkomes et al., 2016). These approaches are often effective, but run the risk of overfitting due to the large number of parameters they introduce. Morever, posing a flexible parametric form for the kernel or spectrum typically presents an optimisation problem which is riddled with local optima. Other approaches treat the kernel as a latent function by assuming a prior distribution over the kernel (Tobar et al., 2015a; Oliva et al., 2016; Jang et al., 2017). This induces a prior distribution over the PSD—see Figure 1 for an illustration—which is appealing, because it brings the benefits of Bayesian nonparametrics to the spectrum: as more data are observed, more spectral structure of the data is revealed, and the posterior over the spectrum automatically increases in complexity. Inference in these models, however, is considerably more involved and computationally demanding.

An example of a model that treats the kernel as a random function is the Gaussian Process Convolution Model (GPCM) (Tobar et al., 2015a), which is the focus of this paper. The GPCM parametrises the kernel of the Gaussian process f with a sample of *another* Gaussian process h, where in turn h has a modified EQ kernel. Tobar et al. (2015b) construct the GPCM by considering a linear system excited by white noise; this construction will be central in this paper.

Although the GPCM works well on a range of tasks, a consequence of the form of the kernel for h is that f has a rapidly decaying spectrum (see Fig 1), which means that f will be a smooth function. If data are not (noisy) observations of a smooth function, the GPCM can fail to capture important structure and lead to predictions which are too smooth and consequently overshoot the data. In addition, the existing inference procedure by Tobar et al. (2015a) relies on a mean-field assumption and is computationally expensive because it requires numerical optimisation over high-dimensional covariance matrices.

The purpose of this paper is twofold: to redesign the GPCM for non-smooth signals and to improve inference in terms of both approximation quality and computational expense. Our contributions are as follows. First, we propose two variations of the GPCM which induce a richer distribution over the spectrum with relaxed assumptions about smoothness: the Causal Gaussian Process Convolution Model (CGPCM) introduces a causality assumption into the GPCM, and the Rough Gaussian Process Convolution Model (RGPCM) can be interpreted as a non-parametric generalisation of the fractional Ornstein–Uhlenbeck process. Second, we propose an improved variational inference scheme which goes beyond the mean-field assumption. In particular, we design a Gibbs sampler which directly samples from the optimal variational solution, which entirely circumvents any variational optimisation; the Gibbs sampler is found to mix quickly and give uncertainty estimates superior to approaches that apply explicit optimisation. Finally, we validate the proposed variations of the GPCM and inference scheme in experiments on synthetic and real-world data.

2 GAUSSIAN PROCESS CONVOLUTION MODELS

The GPCM admits two equivalent formulations. The first formulation of the GPCM is a linear system excited by white noise with a nonparametric prior over the filter (Tobar et al., 2015a). This formulation is useful because it shows how the GPCM is constructed and hence how it can be modified to adjust properties. It also forms the basis for an approximate inference scheme. The second formulation of the GPCM is as a GP with a nonparametric prior over the kernel, which is the interpretation that we are ultimately after.

Let k_h be a kernel with finite trace, meaning that $\int_{-\infty}^{\infty} k_h(\tau, \tau)\, \mathrm{d}\tau < \infty$.[1] Then the linear system formu-

[1]The kernel k_h is then said to be a trace class Hilbert–Schmidt kernel (Lax, 2002).

lation of the GPCM is given by the following model:

$$x \sim \mathcal{GP}(0, \delta(t-t')), \quad h \sim \mathcal{GP}(0, k_h(t,t')), \tag{1}$$
$$f(t) \,|\, h, x = \textstyle\int_{-\infty}^{\infty} h(t-\tau)x(\tau)\,\mathrm{d}\tau, \tag{2}$$

where $\delta(\,\cdot\,)$ denotes the Dirac delta function. This model is accompanied by the data likelihood $y \,|\, f \sim \mathcal{GP}(f(t), \sigma^2\delta[t-t'])$, where $\delta[\,\cdot\,]$ is the Kronecker[2] delta function. From the requirement that k_h has a finite trace it follows that f has finite power: $\mathbb{V}[f(t)] = \int_{-\infty}^{\infty} k_h(\tau,\tau)\,\mathrm{d}\tau < \infty$. One family of kernels with finite trace is given by the *amplitude-modulated, locally stationary* (AMLS) kernels (Chen, 2018) $k_h(t,t') = w(t)w(t')k_g(t-t')$ where w is square integrable and k_g a stationary kernel. Indeed, then $\mathbb{V}[f(t)] = k_g(0)\int_0 w^2(\tau)\,\mathrm{d}\tau < \infty$. Choosing k_h to be an AMLS kernel corresponds to a generative model where we first draw a filter g and then apply a window function w to obtain h: $g \sim \mathcal{GP}(0, k_g)$ and $h(t) \,|\, g = w(t)g(t)$. The window w also serves to make inference well posed: since x is stationary, any shifted version of h results in an identical model for $f \,|\, h$, but versions of h contained within the window are preferred. Following Tobar et al. (2015a), we choose $w(t) = e^{-\alpha t^2}$ and $k_g(t-t') = e^{-\gamma(t-t')^2}$. With these choices, $k_h(t,t') = e^{-\alpha t^2 - \alpha t'^2 - \gamma(t-t')^2}$, which Tobar et al. name the *decaying exponentiated quadratic* (DEQ) kernel. For the DEQ kernel, α determines the temporal extent of the filter h and γ the time scale on which the filter h varies.

When conditioned on h, f is a fixed linear transform (h is fixed) of the Gaussian process x. Consequently, $f \,|\, h$ is also a Gaussian process, with zero mean, $\mathbb{E}[f(t)|h] = 0$, and covariance $k_{f\,|\,h}(t,t') = \mathbb{E}[f(t)f(t')|h]$. This reveals an equivalent formulation of the GPCM with a nonparametric prior over the kernel:

Model 2.1 (GPCM)**.** Let k_h be a DEQ kernel. Then the GPCM is given by the following generative model:

$$h \sim \mathcal{GP}(0, k_h(t,t')), \tag{3}$$
$$f \,|\, h \sim \mathcal{GP}(0, \textstyle\int_{-\infty}^{\infty} h((t-t')+\tau)h(\tau)\,\mathrm{d}\tau). \tag{4}$$

Observe that, in Mod 2.1, the parametrisation of the kernel $k_{f\,|\,h}$ of f is precisely the functional analogue of the parametrisation of a covariance matrix with the outer product: $\mathbf{\Sigma} = \mathbf{A}\mathbf{A}^\top$. As alluded to in the introduction, a consequence of the strong smoothness assumptions on h, which derive from the DEQ kernel, is that the GPCM also exhibits smoothness:

Proposition 2.1. Sample paths of the GPCM are almost surely everywhere differentiable. See App A.

[2] $\delta[0] = 1$ and $\delta[\,\cdot\,] = 0$ elsewhere.

The Causal GPCM. The linear system formulation of the GPCM in (2) is an *acausal* system, meaning that past system responses can depend on future inputs.[3] Acausality combined with the smoothness of h lies at the heart of the smoothness of the GPCM. Therefore, to build a model which is less smooth, we adjust the convolution in (2) to be *causal*. In a causal system, a system response can only depend on past inputs, not future inputs, which is in line with physical systems. Following the construction of the GPCM gives rise to the *Causal GPCM* (see App C):

Model 2.2 (CGPCM)**.** Let k_h be a DEQ kernel. The CGPCM is given by the following generative model:

$$h \sim \mathcal{GP}(0, k_h(t,t')), \tag{5}$$
$$f \,|\, h \sim \mathcal{GP}(0, \textstyle\int_0^{\infty} h(|t-t'|+\tau)h(\tau)\,\mathrm{d}\tau). \tag{6}$$

The only difference between Mod 2.2 and Mod 2.1 is that the integral starts at zero and depends on $|t-t'|$ rather than on $t-t'$. As the next proposition shows, this seemingly minor detail has major consequences for the smoothness properties of the CGPCM.

Proposition 2.2. If $h(0) = 0$, then sample paths of the CGPCM are almost surely everywhere differentiable. If, on the other hand, $h(0) \neq 0$, then sample paths of the CGPCM are almost surely nowhere differentiable. See App B for a proof.

Although Prop 2.2 tells us that sample paths of the CGPCM are almost surely nowhere differentiable in the case $|h(0)| > 0$, the proposition does not give us a sense of how volatile the sample paths then are. App C argues f that can locally be approximated by a $|h(0)|$-scaled Brownian motion, which means that the magnitude of the irregular increments is controlled by the value of $|h(0)|$. Intuitively, $|h(0)|$ is the magnitude of the filter when new white noise enters it: if $|h(0)| > 0$, new noise is directly passed to the output, which results in a nondifferentiable signal; and the larger $|h(0)|$ is, the more noisy the output will be. Fig 2 demonstrates this mechanism. Since h is modelled randomly, by performing inference in the CGPCM, the model is able to automatically infer a level of irregularity, *i.e.* a value for $|h(0)|$, which is appropriate for the data.

The Rough GPCM. The CGPCM is able to model nondifferentiable phenomena. This is visualised by samples in the bottom row of Fig 2, which look fairly jagged. Some applications, however, may require samples which behave even more erratically, like equilibrium systems under noise in the natural sciences, and certain financial time series. To this end, we modify

[3] Although similarly named, the system-theoretic notion of causality here is distinct from the probabilistic notion of causality.

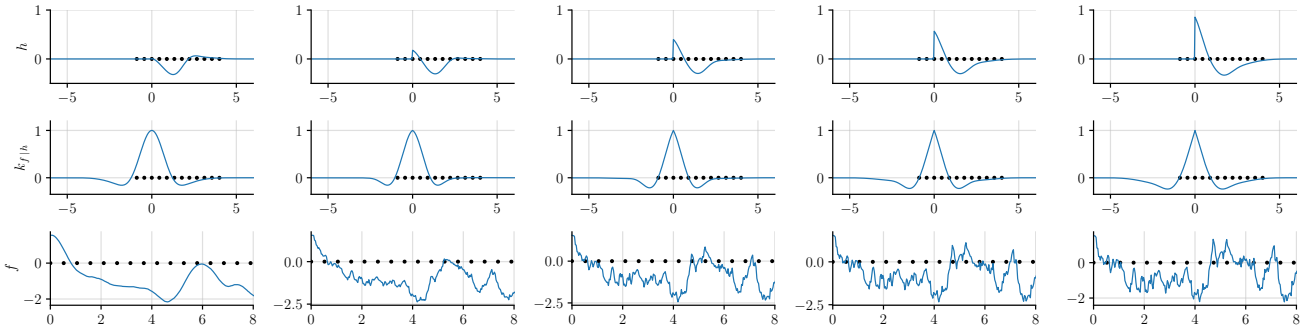


Figure 2: Generative process of the CGPCM. Shows the filter h, the kernel $k_{f\,|\,h}$, and a sample $f\,|\,h \sim \mathcal{GP}(0, k_{f\,|\,h}(t-t'))$ while the filter is interpolated from one that satisfies $h(0) = 0$ to one that satisfies $|h(0)| > 0$. The sample appears smooth for $h(0) = 0$ and becomes more irregular as $|h(0)|$ increases. The black dots indicate inducing point positions.

both the filter h and input x of the (C)GPCM: we relax the smoothness of the filter h for greater spectral flexibility, and posit a Matérn–$\frac{1}{2}$ kernel for the input process x. As will be explored in the next section, instead of using a smoothing transformation for inducing points for x, which the (C)GPCM use, our construction will enable efficient inference of spectral content through variational Fourier features. The RGPCM can be interpreted as altering a Matérn–$\frac{1}{2}$ GP—also known as an Ornstein–Uhlenbeck (OU) process—by a random nonparametric modulation of the spectrum. A parametric special case of this model is the fractional Ornstein–Uhlenbeck process (Cheridito et al., 2003), which can be *rough* in the sense that it is more irregular than the OU process.[4] We therefore call this version the *Rough GPCM*.

Model 2.3 (RGPCM)**.** Let h be white noise windowed by $w(t) = e^{-\alpha|t|}$ and $k_x(t,t') = e^{-\lambda|t-t'|}$. Then the RGPCM is given by the following model:

$$h \sim \mathcal{GP}(0, k_h(t,t')), \quad f\,|\,h \sim \mathcal{GP}\left(0, k_{f|h}(t-t')\right) \quad (7)$$

with $k_{f\,|\,h}(r) = \int_0^\infty \int_0^\infty h(\tau)h(\tau')k_x(r-(\tau-\tau'))\,\mathrm{d}\tau\,\mathrm{d}\tau'$.

See App D for a more detailed description of the RGPCM. As we will see next, the RGPCM allows for more spectral content over higher frequencies and thus more irregular sample paths.

Comparison of model priors. Fig 3 visualises the nonparametric prior distribution over kernels and PSDs induced by the GPCM, CGPCM, and RGPCM. Observe that the GPCM has a very quickly decaying spectrum, whereas the CGPCM and especially the RGPCM have substantial support at higher frequencies. This is in line with the construction of the models: the GPCM models smooth signals, the CGPCM models signals with varying levels of irregularity, and the RGPCM models the most irregular signals. To further support this, Fig 4 shows function, kernel, and PSD samples. Crucially, observe that GPCM samples are smooth, the CGPCM varies in level of smoothness (*e.g.*, the green sample is smooth and yellow one is jagged), and the RGPCM is very irregular.

Choice of hyperparameters. To fairly compare the GPCM, CGPCM, and RGPCM in experiments, we need to able to configure the models comparably. By requiring the models' prior powers to be unity and that, for an appropriate definition of the length scale, the prior marginal covariance functions $\mathbb{E}[k_{f\,|\,h}(t,t')]$ have equal length scales, App E derives the following initialisation. For some data, let τ_f be the smallest length scale contained in the signal and let τ_w be the desired extent of the filter. Let the subscript $\cdot_{\mathrm{c}}$ refer to the CGPCM, $\cdot_{\mathrm{ac}}$ to the GPCM, and $\cdot_{\mathrm{r}}$ to the RGPCM. Then initialise $\alpha_{\mathrm{ac,c}} = \frac{\pi}{4}\tau_w^{-2}$, $\alpha_{\mathrm{r}} = \tau_w^{-1}$, $\gamma_{\mathrm{ac,c}} = \frac{\pi}{4}\tau_f^{-2} - \frac{1}{2}\alpha_{\mathrm{ac,c}}$, and $\lambda_{\mathrm{r}} = \tau_f^{-1}$. By setting $\tau_w = 2\tau_f$, $\tau_{f,\mathrm{ac,c}} = \sqrt{\pi/2}\ell \approx 1.2\ell$, and $\tau_{f,\mathrm{rv}} = \ell$, we define the *standardised marginal covariance functions*:

$$k_{\mathrm{ac}}(r) = \exp(-\tfrac{1}{2\ell^2}r^2), \quad (8)$$

$$k_{\mathrm{c}}(r) = (1 - \mathrm{erf}(\tfrac{1}{4\ell}|r|))\exp(-\tfrac{1}{2\ell^2}r^2), \quad (9)$$

$$k_{\mathrm{r}}(r) = \exp(-\tfrac{1}{\ell}|r|). \quad (10)$$

To the best of our knowledge, the kernel k_{c} does not have an established name. (That k_{c} is a positive definite function follows from the fact that it is a covariance function of the CGPCM.) We call k_{c} the *causal exponentiated quadratic* (CEQ) kernel. The standard kernels (8) to (10) are helpful, because they allow us to build intuition for the GPCM models by comparing to familiar kernels: the GPCM is like an EQ GP, the CGPCM is also like an EQ GP but with an irregular component, and the RGPCM is like a Matérn–$\frac{1}{2}$ GP.

3 INFERENCE

For all GPCM models, the posterior conditioned on data cannot be computed analytically, so an approx-

[4] Formally, the OU process is Hölder continuous of order $a < \frac{1}{2}$ while the fractional OU has $a < H$ for $H \in (0,1)$. If $H < \frac{1}{2}$, then it is called rough; see Gatheral et al. (2018) and Bennedsen et al. (2016) for empirical evidence of roughness in financial data.

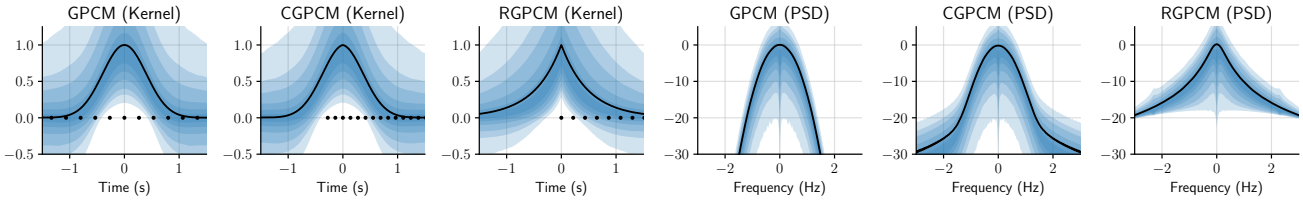


Figure 3: Visualisation of the nonparametric priors induced over kernels and PSDs by the GPCM, CGPCM, and RGPCM with $\tau_f = 0.5$ s and $\tau_w = 2$ s (see Sec 3). The black line shows the mean and the shaded areas show marginal quantiles ranging from 1% to 99%. The number of inducing points is $n_u = 30$; the black dots indicate inducing point positions.

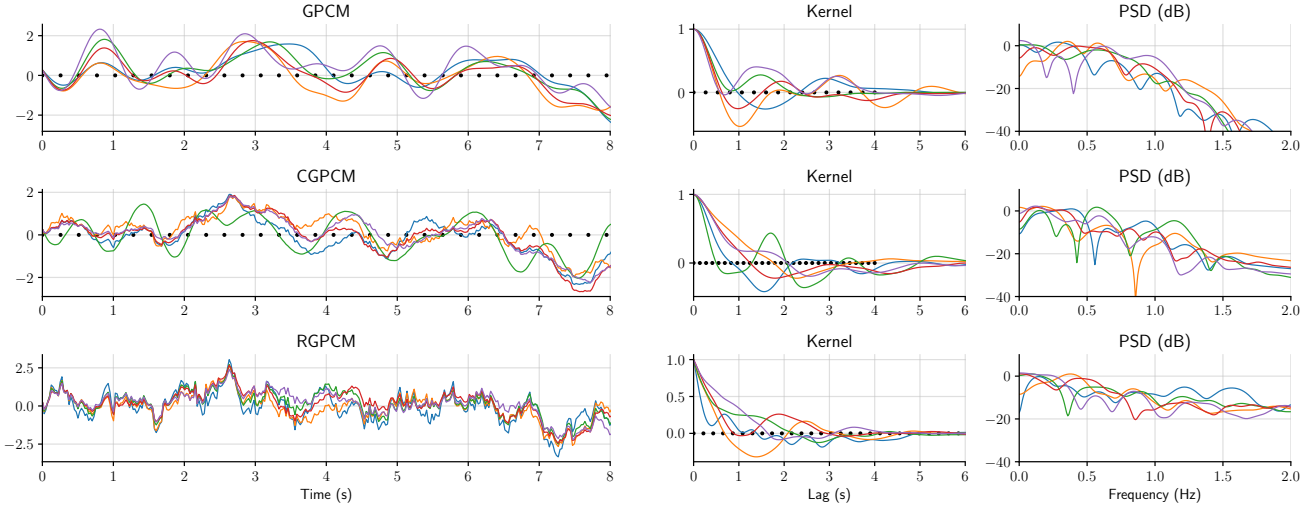


Figure 4: Prior function, kernel, and PSD samples from the GPCM, CGPCM, and RGPCM with $\tau_f = 0.5$ s and $\tau_w = 2$ s (see Sec 3). The numbers of inducing points are $n_u = 30$ and $n_z = 40$; black dots indicate inducing point positions.

imation is necessary. We follow Tobar et al. (2015a) and consider a variational approximation (Wainwright and Jordan, 2008). The setup and derivation of our inference scheme is spelled out in detail in App F; in this section, we give a high-level sketch.

To approximate the posterior over the Gaussian processes x and h, we make use of inducing points (Titsias, 2009; d. G. Matthews et al., 2016). For the GPCM and CGPCM, let $\mathbf{u}$ be n_u inducing points for h, and let $\mathbf{z}$ be n_z inducing points for the inter-domain transform $s(\tau) = \int_{-\infty}^{\infty} e^{-\omega(\tau-t)^2} x(t)\,\mathrm{d}t$ (Lázaro-Gredilla and Vidal, 2009). Note that s has an EQ kernel. For the RGPCM, first, let $\mathbf{u}$ be n_u inducing points for the *causal* inter-domain transform $s(\tau) = \int_{-\infty}^{\tau} e^{-\gamma(\tau-t)} x(t)\,\mathrm{d}t$. Compared to the GPCM, s now has a Matérn–$\frac{1}{2}$ kernel. Second, for h, we exploit the fact that x also has a Matérn–$\frac{1}{2}$ kernel, which enables us to *variational Fourier features* (VFFs) (Hensman et al., 2018). Whereas regular inducing points approximate the posterior with temporally local basis functions placed at the inducing points, VFFs approximate the posterior with a truncated Fourier series, which can give superior spectral approximation qualities. Because x is stationary, such an approximation is not possible with the Fourier transform as inter-domain transform; rather, VFFs propose a clever construction which exploits the fact that harmonics are contained within the reproducing kernel Hilbert space of the Matérn–$\frac{1}{2}$ kernel. See Hensman et al. (2018) for more details. For the RGPCM, we let $\mathbf{z}$ be n_z VFFs for x.

Henceforth, let θ denote all hyperparameters of a model. Given the inducing points $\mathbf{u}$ and $\mathbf{z}$, we follow Tobar et al. and consider the variational approximation $q_\theta(h, x, \mathbf{u}, \mathbf{z}) = p_\theta(h \mid \mathbf{u}) p_\theta(x \mid \mathbf{z}) q(\mathbf{u}, \mathbf{z})$ where $q(\mathbf{z}, \mathbf{u})$ is a joint variational approximation over the inducing points. Given some data $\mathbf{y}$, to optimise $q(\mathbf{z}, \mathbf{u})$ and θ, we optimise the *evidence lower bound* (ELBO):

$$\begin{aligned} &\mathcal{F}_\theta[q(\mathbf{u}, \mathbf{z})] \qquad (11)\\ &= \mathbb{E}_q[\log p_\theta(\mathbf{y} \mid f)] - \mathrm{KL}[q(\mathbf{u}, \mathbf{z}) \,\|\, p_\theta(\mathbf{u}) p_\theta(\mathbf{z})]. \end{aligned}$$

As we explain in App F, key to the tractability of the ELBO is the observation that $\mathbb{E}[\log p_\theta(\mathbf{y} \mid f) \mid \mathbf{u}, \mathbf{z}]$ is tractable and conditionally quadratic in $\mathbf{u}$ and $\mathbf{z}$.

Mean-field inference. The first scheme that we consider is the *mean-field* (MF) *approximation* $q(\mathbf{u}, \mathbf{z}) = q(\mathbf{u}) q(\mathbf{z})$, originally considered by Tobar et al. (2015a). They parametrise $q(\mathbf{u})$ and $q(\mathbf{z})$ by Gaussians with dense covariance matrices and optimise the ELBO using gradient-based optimisation. In App F, we show that, given $q(\mathbf{z})$ (resp. $q(\mathbf{u})$), the optimal $q^*(\mathbf{u})$ (resp. $q^*(\mathbf{z})$) can be computed analytically, which gives rise to a coordinate ascent (CA) scheme. Alternatively, the

optimal form $q^*(\mathbf{z})$ (resp. $q^*(\mathbf{u})$) can be plugged back into the ELBO to give rise a collapsed MF bound, which depends on many fewer variational parameters and hence greatly accelerates optimisation.

Structured inference. A major issue with the mean-field approximations is that it is unable to model correlations between $\mathbf{u}$ and $\mathbf{z}$. It consequently biases towards overly simple models and and tends to yield poorly calibrated uncertainties (MacKay, 2002; Turner and Sahani, 2011). To improve upon the mean-field approximation, we consider a general *structured approximation* $q(\mathbf{u}, \mathbf{z})$. In App F, we derive the optimal $q^*(\mathbf{u}, \mathbf{z})$ and demonstrate that the conditionals $q^*(\mathbf{u} \,|\, \mathbf{z})$ and $q^*(\mathbf{z} \,|\, \mathbf{u})$ are Gaussians with parameters dependent on respectively $\mathbf{z}$ and $\mathbf{u}$. Therefore, to sample from the optimal $q^*(\mathbf{u}, \mathbf{z})$, we can iteratively sample from these conditionals in an alternating fashion. This gives us a way to perform inference in the models without any variational optimisation, and which even enjoys computational benefits compared to the mean-field schemes (see App F). To optimise θ, App F shows that samples from $q^*(\mathbf{u}, \mathbf{z})$ can be used to approximate $\frac{\mathrm{d}}{\mathrm{d}\theta}\mathcal{F}_\theta[q^*(\mathbf{u}, \mathbf{z})]$. Although gradients can be approximated, $\mathcal{F}_\theta[q^*(\mathbf{u}, \mathbf{z})]$ unfortunately cannot be estimated. As a proxy, we propose the lower bound $\mathcal{F}_\theta[q^*_{\mathrm{MF}}(\mathbf{u})q^*(\mathbf{z} \,|\, \mathbf{u})] \leq \mathcal{F}_\theta[q^*(\mathbf{u}, \mathbf{z})]$, which can be estimated. Here $q^*_{\mathrm{MF}}(\mathbf{u})$ is the optimal MF solution.

4 EXPERIMENTS

We provide an implementation of the GPCM, CGPCM, and RGPCM at `github.com/wesselb/gpcm` with a user-friendly `sklearn`-style interface. In this section, we validate the proposed variants of the GPCM and inference scheme on synthetic and real-world data. We use mean log loss (MLL)[5] as the metric to evaluate uncertainty and the root-mean-square error (RSME) as the metric to evaluate accuracy of the mean prediction. Unlike Tobar et al. (2015a), in all experiments we optimise the inducing point locations and all hyperparameters.

Learning performance of inference schemes. We compare the inference schemes described in Sec 3 to the original inference scheme by Tobar et al. (2015a). Fig 5 shows the evolution of the ELBO over wall-clock time for all inference schemes in a toy problem. Note that the collapsed MF bound optimises quicker than the uncollapsed MF bound, but that the coordinate ascent procedure (CA) converges nearly instantaneously and reaches its convergence threshold long before the gradient-based optimisation. Moreover, this example

[5]For predictions given by means and marginal variances, the mean log loss is the average negative log-pdf of the observations under those means and marginal variances.

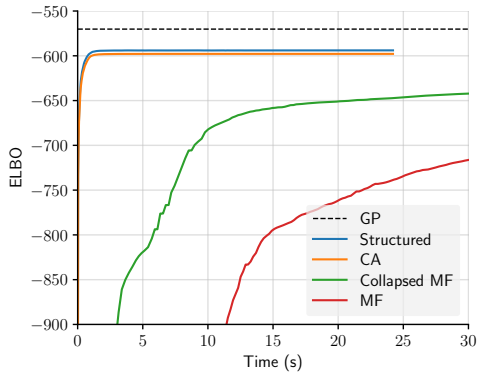
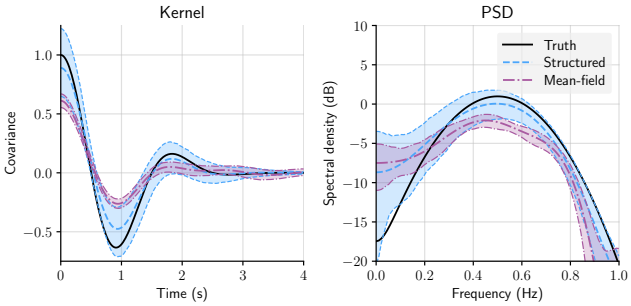


Figure 5: Learning 500 noisy data points sampled from a GP with an EQ kernel. Shows the evolution of the ELBO in time for the original mean-field approximation by Tobar et al. (2015a) (MF), a collapsed version of the mean-field approximation (collapsed MF), the coordinate-ascent version of the mean-field approximation (CA), and the structured ELBO with $q(\mathbf{u})$ given by the current CA solution; see Sec 3. Also shows the likelihood of the GP from which the data was sampled. `scipy`'s implementation of the L-BFGS-B algorithm (Nocedal and Wright, 2006) was used to optimise the uncollapsed and collapsed mean-field ELBO. The numbers of inducing points are $n_u = 40$ and $n_z = 40$.

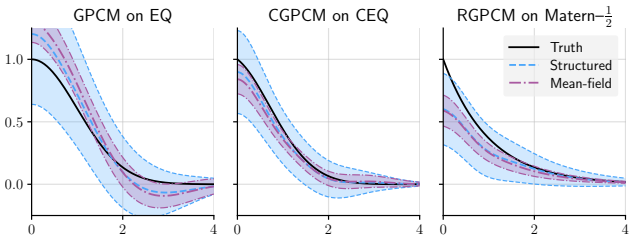
only uses 40 inducing points. The benefits of the CA scheme will be further exaggerated for greater numbers of inducing points, because gradient-based optimisation will then struggle with the large covariance matrices. Finally, observe that the structured scheme, which derives from the CA solution (see Sec 3), further improves the ELBO and comes closest to the GP likelihood out of all inference schemes.

Approximation quality of inference schemes. In this experiment, a reference to the mean-field approximation scheme will refer to the CA scheme, followed by optimisation of the hyperparameters with the collapsed MF ELBO, followed by another application of the CA scheme (see Sec 3). Fig 6 presents the results when we use the GPCM to infer the kernel and PSD of a GP with a one-component spectral mixture kernel (Wilson and Adams, 2013) from a noisy sample. Observe that the mean-field scheme produces a poor solution for the predictive mean and that the uncertainties are uncalibrated; in contrast, the structured scheme is able to capture true kernel and PSD. We further compare the mean-field and structured approximation in a second experiment. Fig 7 presents the results of using the GPCM, CGPCM, and RGPCM to infer their respective standard kernels from a noisy sample. Observe that, in all cases, the structured scheme presents marked improvements in MLL; indeed, Fig 7 shows that the true kernels hit the edges or even just fall outside of the uncertainty intervals produced by the mean-field scheme. In this case, however, the structured scheme did not yield improvements in RMSE. Generally, the benefit of the structured scheme consti-



Model	Data	MLL		RMSE	
		MF	S	MF	S
GPCM	SMK	35.8	**−1.39**	0.19	**0.07**

Figure 6: Fitting the GPCM on 200 noisy observations drawn from a GP with a one-component spectral mixture kernel using the mean-field inference scheme (MF) and structured inference scheme (S). The numbers of inducing points are $n_u = 80$ and $n_z = 80$. Shows the prediction for the kernel and for the PSD. Also shows the MLL and RMSE of the kernel prediction for both inference schemes. Best numbers are boldfaced.



Model	Data	MLL		RMSE	
		MF	S	MF	S
GPCM	EQ	−1.35	**−1.44**	0.10	**0.08**
CGPCM	CEQ	−1.23	**−1.73**	**0.09**	0.12
RGPCM	Mat−$\frac{1}{2}$	−0.88	**−1.59**	**0.05**	**0.05**

Figure 7: Fitting the GPCM, CGPCM, and RGPCM on 400 noisy observations drawn from a GP with respectively a CEQ (see Sec 3), EQ, and Matérn−$\frac{1}{2}$ kernel using the mean-field inference scheme (MF) and structured inference scheme (S). The numbers of inducing points are $n_u = 30$ and $n_z = 80$. Shows the prediction for the kernel and the MLL and RMSE for both inference schemes. Numbers within 1% of the best number are boldfaced.

tutes improved uncertainty estimates. Having demonstrated the advantages of the structured scheme, we commit to the structured scheme in the remaining experiments with real data.

Predicting crude oil prices. We demonstrate the benefits of the relaxed smoothness assumptions of the CGPCM and RGPCM by, for the years 2012–2017, predicting NASDAQ crude oil daily prices[6] in every odd week of the second half of the year from all other data points in that year. Fig 8 presents the results. To begin with, we focus on the predictions by the models in the top two plots. Observe that the predictions by the GPCM (blue) are much smoother than the predictions by the CGPCM (purple) and RGPCM (green). The smoothness of the GPCM's predictions causes the model to explain more intricate structure of the signal as noise, which consequently leads to a significant increase in MLL and RMSE. That the predictions of the GPCM are too smooth is corroborated by the predictions for the PSD: the GPCM predicts a quickly decaying EQ-like spectrum, whereas the CGPCM and RGPCM predict support at higher frequencies and exhibit more fine-grained spectral structure.

Forecasting the Cboe Volatility Index. The Cboe Volatility Index[7] (VIX) is an index which measures the market's expectations for short-term S&P500 price changes. In this experiment, we train the models on a randomly chosen year between 1990 and 2010, retain the posterior over h, move forward a year and,

[6] `https://www.nasdaq.com/market-activity/commodities/cl%3Anmx`

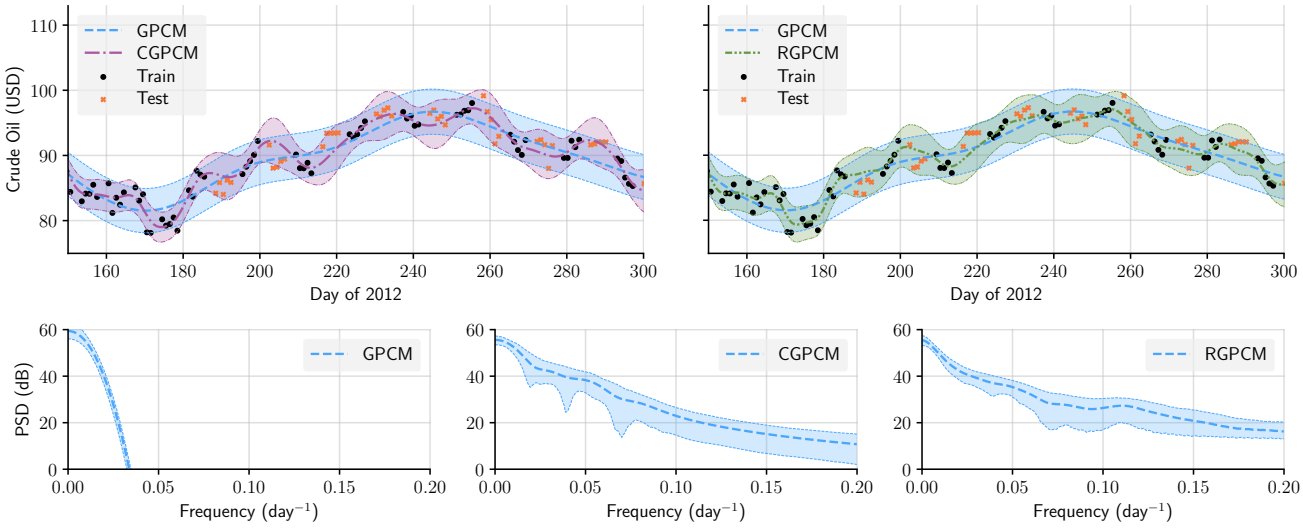
[7] `https://www.cboe.com/tradable_products/vix/vix_historical_data/`

	GPCM	CGPCM	RGPCM
MLL	−0.497 $\pm$0.126	−0.543* $\pm$0.096	**−0.681*** $\pm$0.100
RMSE	0.107* $\pm$0.005	0.112 $\pm$0.005	**0.101*** $\pm$0.005

Table 1: Average one-week-ahead prediction result for log-VIX (see Sec 4). Shows the MLL and RSME for the GPCM, CGPCM, and RGPCM. Best numbers boldfaced. *Significantly better than all worse scores with $p < 10^{-4}$ using a paired test.

in a one-week rolling window fashion, for 100 weeks, predict log-VIX one week ahead given the past four weeks. Table 1 presents the results. Whereas the CGPCM outperforms the GPCM in terms of MLL, which means that the causality assumption is helpful, the even weaker smoothness assumptions of the RGPCM yield the best uncertainty and mean estimates.

Analysing the Cboe Volatility Index. In the final experiment, we use the RGPCM to investigate on which length scale the log-VIX reasonably be approximated by an OU process. We fit the RGPCM to all log-VIX in the year 2000 using a number of inducing points which allows the RGPCM to detect fluctuations up to the Nyquist frequency. We then use the property of the RGPCM that it is a spectrally modulated version of an OU process (see Sec 3 and App D): the prediction of the PSD $|\mathcal{F}h(f)|^2 \mathcal{F}k_x(f)$ by the RGPCM, where $\mathcal{F}$ denotes the Fourier transformation, can be decomposed into the prediction of the spectrum $\mathcal{F}k_x(f)$ for the OU process x and a prediction of the modulation by the filter $|\mathcal{F}h(f)|^2$. Fig 9 shows that the prediction for $|\mathcal{F}h(f)|^2$ is flat after $f = 0.2$ day^{-1}. We conclude that the log-VIX data can reasonably be modelled with an OU process on a length scale



	GPCM	CGPCM	RGPCM
MLL	-1.803 ± 0.060	$\mathbf{-1.680}^{*} \pm 0.072$	$\mathbf{-1.665}^{*} \pm 0.071$

	GPCM	CGPCM	RGPCM
RMSE	1.459 ± 0.082	$\mathbf{1.330}^{*} \pm 0.083$	$\mathbf{1.302}^{*} \pm 0.067$

Figure 8: Predictions for NASDAQ crude oil prices by the GPCM, CGPCM, RGPCM for every odd week in the second half of 2012 (see Sec 4). Tables show the MLL and RMSE of these predictions. Also shows the predictions for the PSDs. Models have $n_u = 50$ and $n_z = 150$. Best numbers are boldfaced. *Significantly better than the GPCM with $p < 0.05$ using a paired test.

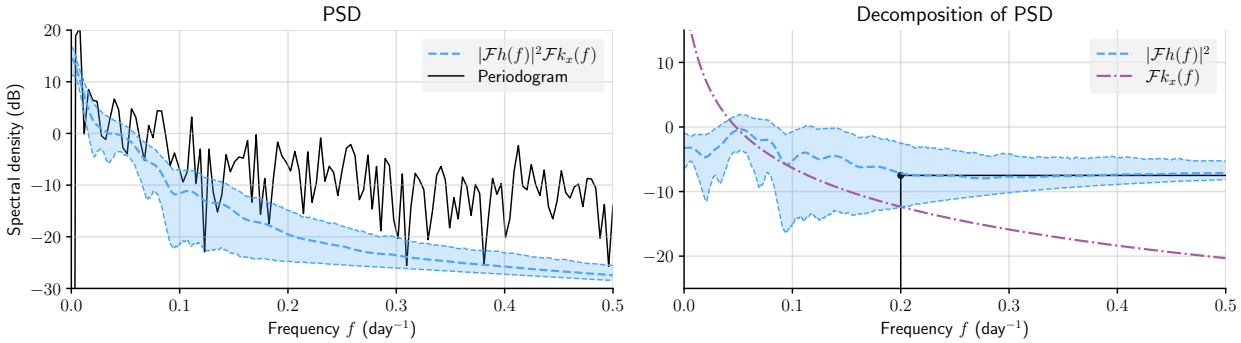


Figure 9: Prediction for the PSD of log-VIX from the year 2000 by the RGPCM (see Sec 4). Shows the periodogram, the whole prediction $|\mathcal{F}h(f)|^2\mathcal{F}k_x(f)$ for the PSD, the Matérn–$\frac{1}{2}$ part $\mathcal{F}k_x(f)$ of the predicted PSD, and the modulation by the filter $|\mathcal{F}h(f)|^2$. The numbers of inducing points are $n_u = 60$ and $n_z = 365$, which allows the RGPCM to detect fluctuations up to the Nyquist frequency. Observe that the prediction for $|\mathcal{F}h(f)|^2$ is flat after $f = 0.2$ day^{-1}.

of at most $0.2^{-1} = 5$ days; on longer length scales, the RGPCM predicts more intricate spectral structure.

5 RELATED WORK

Tobar et al. (2015b) extend the GPCM by considering an harmonic inter-domain transformation for x, making the model more suitable for more complex spectral estimation tasks. In the supplement, Tobar et al. (2015a) point out that the PSD of $f \mid \mathbf{u}$ is a mixture of Gaussians centred at frequencies determined by the inducing point locations. From this point of view, the GPCM is related to the models by Oliva et al. (2016), who use a Dirichlet process to parametrise the PSD; by Jang et al. (2017), who use a Levy process for the kernel, resulting in PSDs consisting of Laplacian mixtures; and by Benton et al. (2019), who model the log-density of the PSD with a GP. However, to perform inference, all these models employ general-purpose MCMC procedures; in contrast, our inference scheme exploits the additional structure that the conditionals of the optimal variational solution can be sampled from directly to construct a Gibbs sampler which mixes quickly in practice. Finally, a construction like the GPCM can also be found in other fields: Pillonetto and Nicolao (2010); Wågberg et al. (2018); Chen (2018) use a frequentist approach similar to GPCM for the purpose of system identification.

6 DISCUSSION

The goal of this paper was to redesign the GPCM to induce a richer distribution over the spectrum. We introduced the *Causal GPCM*, able to model signals of varying level of irregularity, and the *Rough GPCM*, able to model even more irregular signals. The RGPCM is particularly appealing, because it avoids the implementation difficulties of the CGPCM and enjoys the benefits of variational Fourier features for improved approximation qualities. Experiments demonstrated that the relaxed smoothness assumptions of the CGPCM and RGPCM can yield substantially improved uncertainty and mean estimates on real-world data. In addition, to address the deficiencies of the original mean-field inference scheme by Tobar et al. (2015a), we introduced a structured approximation which is able to fully retain the correlation structure between the latent variables. Experiments showed that the structured scheme generally gives marked improvements in uncertainty estimates and can perform well in cases where the mean-field scheme falls over.

Acknowledgements

We thank anonymous referees for helpful comments and discussions. Wessel P. Bruinsma was supported by the Engineering and Physical Research Council (studentship number 10436152). Martin Tegnér thanks the Oxford-Man Institute for research support. Richard E. Turner is supported by Google, Amazon, ARM, Improbable and EPSRC grant EP/T005386/1.

References

Felipe Tobar, Thang D. Bui, and Richard E. Turner. Learning stationary time series using Gaussian processes with nonparametric kernels. *Advances in Neural Information Processing Systems*, 29:3501–3509, 2015a.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Thang Bui, Daniel Hernandez-Lobato, Jose Hernandez-Lobato, Yingzhen Li, and Richard Turner. Deep Gaussian processes for regression using approximate expectation propagation. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1472–1481. PMLR, 2016. URL `https://proceedings.mlr.press/v48/bui16.html`.

Stamatis Cambanis. On some continuity and differentiability properties of paths of Gaussian processes. *Journal of Multivariate Analysis*, 3(4):420–434, 1973.

Andrew Wilson and Ryan Adams. Gaussian process kernels for pattern discovery and extrapolation. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1067–1075. PMLR, 2013. URL `https://proceedings.mlr.press/v28/wilson13.html`.

R. Calandra, J. Peters, C. E. Rasmussen, and M. P. Deisenroth. Manifold Gaussian processes for regression. In *International Joint Conference on Neural Networks 2016*, pages 3338–3345, 2016. doi: 10.1109/IJCNN.2016.7727626.

S. Sun, G. Zhang, C. Wang, W. Zeng, J. Li, and R. Grosse. Differentiable compositional kernel learning for Gaussian processes. *International Conference on Machine Learning*, 35, 2018.

Roger Grosse, Ruslan R Salakhutdinov, William T. Freeman, and Joshua B. Tenenbaum. Exploiting compositionality to explore a large space of model structures. In *28th Conference on Uncertainty in Artificial Intelligence*, 2012.

David Duvenaud. *Automatic Model Construction With Gaussian Processes*. PhD thesis, Computational and Biological Learning Laboratory, University of Cambridge, 2014.

Gustavo Malkomes, Charles Schaff, and Roman Garnett. Bayesian optimization for automated model selection. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL `https://proceedings.neurips.cc/paper/2016/file/3bbfdde8842a5c44a0323518eec97cbe-Paper.pdf`.

Junier B. Oliva, Avinava Dubey, Andrew G. Wilson, Barnabas Poczos, Jeff Schneider, and Eric P. Xing. Bayesian nonparametric kernel-learning. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 1078–1086. PMLR, 2016. URL `https://proceedings.mlr.press/v51/oliva16.html`.

Phillip A. Jang, Andrew Loeb, Matthew Davidow, and Andrew G. Wilson. Scalable levy process priors for spectral kernel learning. In *Advances in Neural Information Processing Systems*, volume 30, pages 3940–3949. Curran Associates, Inc., 2017.

Felipe Tobar, Thang D. Bui, and Richard E. Turner. Design of covariance functions using inter-domain inducing variables. In *Advances in Neural Information Processing Systems 28*, 2015b.

Peter D. Lax. *Functional Analysis*. John Wiley & Sons, Inc., 2002.

Tianshi Chen. On kernel design for regularized LTI system identification. *Automatica*, 90:109–122, 2018. ISSN 0005-1098. doi: 10.1016/j.automatica.2017.12.039. URL https://www.sciencedirect.com/science/article/pii/S000510981730626X.

Patrick Cheridito, Hideyuki Kawaguchi, and Makoto Maejima. Fractional Ornstein–Uhlenbeck processes. *Electronic Journal of Probability*, 8:1–14, 2003. doi: 10.1214/EJP.v8-125. URL https://doi.org/10.1214/EJP.v8-125.

Jim Gatheral, Thibault Jaisson, and Mathieu Rosenbaum. Volatility is rough. *Quantitative finance*, 18 (6):933–949, 2018.

Mikkel Bennedsen, Asger Lunde, and Mikko S Pakkanen. Decoupling the short-and long-term behavior of stochastic volatility. *arXiv preprint arXiv:1610.00332*, 2016.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1):1–305, 2008. ISSN 1935-8237. doi: 10.1561/2200000001. URL http://dx.doi.org/10.1561/2200000001.

Michalis K. Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, volume 12 of *Proceedings of Machine Learning Research*, pages 567–574. PMLR, 2009. URL http://proceedings.mlr.press/v5/titsias09a/titsias09a.pdf.

A. G. d. G. Matthews, J. Hensman, R. E. Turner, and Z. Ghahramani. On sparse variational methods and the Kullback-Leibler divergence between stochastic processes. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*. PMLR, 2016.

Miguel Lázaro-Gredilla and Aníbal Figueiras Vidal. Inter-domain Gaussian processes for sparse inference using inducing features. *Advances in Neural Information Processing Systems*, 22:1087–1095, 2009.

James Hensman, Nicolas Durrande, and Arno Solin. Variational fourier features for Gaussian processes. *Journal of Machine Learning Research*, 18(151): 1–52, 2018. URL http://jmlr.org/papers/v18/16-579.html.

David J. C. MacKay. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, 2002.

R. E. Turner and M. Sahani. Two problems with variational expectation maximisation for time-series models. In *Bayesian Time Series Models*, chapter 5, pages 109–130. Cambridge University Press, 2011.

J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 2nd edition, 2006.

Gregory W. Benton, Wesley J. Maddox, Jayson P. Salkey, Julio Albinati, and Andrew Gordon Wilson. Function-space distributions over kernels. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019.

Gianluigi Pillonetto and Giuseppe De Nicolao. A new kernel-based approach for linear system identification. *Automatica*, 46(1):81–93, 2010. ISSN 0005-1098. doi: 10.1016/j.automatica.2009.10.031. URL http://www.sciencedirect.com/science/article/pii/S0005109809004920.

J. Wågberg, D. Zachariah, and T. B. Schön. Regularized parametric system identification: A decision-theoretic formulation. In *Annual American Control Conference*, pages 1895–1900, 2018. doi: 10.23919/ACC.2018.8430895.

Georg Lindgren. *Stationary Stochastic Processes: Theory and Applications*. CRC Press, 2012.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

Supplementary Material: Modelling Non-Smooth Signals with Complex Spectral Structure

Contents

A Proof of Proposition 2.1

Proof of Proposition 2.1. To begin with, h is a sample of a GP with an EQ kernel multiplied by a smooth window w, so h is infinitely differentiable almost surely. Since, in addition, the window $w(t) = e^{-\alpha t^2}$ and all its derivatives decay to zero quickly, almost surely, (1) h and all derivatives of h go to zero at infinity and (2) any product of h and its derivatives is dominated by an integrable function. This is argued more rigorously in App A.1. We will use these fact implicitly in the remainder of the proof to freely interchange integral and derivative.

Note that

$$k_{f\,|\,h}(r) = k_{f\,|\,h}(0) + k'_{f\,|\,h}(0)r + \tfrac{1}{2}k''_{f\,|\,h}(0)r^2 + O(|r|^3), \tag{12}$$

so sample paths of the GPCM are almost surely everywhere differentiable if $k'_{f\,|\,h}(0) = 0$ and $k''_{f\,|\,h}(0) \neq 0$ (Thm 3 from Cambanis, 1973). First, use integration by parts to find

$$k'_{f\,|\,h}(0) = \int_{-\infty}^{\infty} h'(\tau)h(\tau)\,\mathrm{d}\tau = \cancel{\lim_{\tau\to\infty}\left[h^2(\tau) - h^2(-\tau)\right]} - \int_{-\infty}^{\infty} h(\tau)h'(\tau)\,\mathrm{d}\tau = -k'_{f\,|\,h}(0) \tag{13}$$

almost surely. Thus, $k'_{f\,|\,h}(0) = 0$ almost surely. Second, again use integration by parts to find

$$k''_{f\,|\,h}(0) = \int_{-\infty}^{\infty} h''(\tau)h(\tau)\,\mathrm{d}\tau = \cancel{\lim_{\tau\to\infty}\left[h'(\tau)h(\tau) - h'(-\tau)h(-\tau)\right]} - \int_{-\infty}^{\infty} (h'(\tau))^2\,\mathrm{d}\tau < 0 \tag{14}$$

almost surely, so $k''_{f\,|\,h}(0) \neq 0$ almost surely. □

A.1 Well-Behavedness of the Filter

We show that the filter h is well behaved. In particular, we show that, almost surely, (1) h and all derivatives of h go to zero at infinity and (2) any product of h and its derivatives is dominated by an integrable function.

Definition A.1. A function $w\colon \mathbb{R} \to \mathbb{R}$ is said to *decay (sufficiently) quickly* if there exist $p_i \geq 0$ and $\alpha_i, \beta_i > 0$ such that

$$|w(t)| \leq \sum_{i=1}^{n} |t|^{p_i} \exp(-\alpha_i |t|^{\beta_i}) \tag{15}$$

for all $t \in \mathbb{R}$.

Lemma A.1. Let w decay sufficiently quickly, and let $(t_n)_{n\geq 1} \subseteq \mathbb{R}$, $t_n \uparrow \infty$ be such that $t_n \geq n$ for all n. Then $\sum_{n=1}^{\infty} n\,|w(t_n)| < \infty$.

Proof. To begin with, note that

$$\sum_{n=1}^{\infty} n|t_n|^p e^{-\alpha t_n^{\beta}} \leq \sum_{n=1}^{\infty} \lfloor t_n \rfloor (\lfloor t_n \rfloor + 1)^p \exp(-\alpha \lfloor t_n \rfloor^{\beta}). \tag{16}$$

Consider the sequence

$$a_n = n(n+1)^p \exp(-\alpha n^{\beta}) \leq 2^p n^{p+1} \exp(-\alpha n^{\beta}). \tag{17}$$

We claim that $\sum_{n=1}^{\infty} a_n < \infty$, where the convergence is absolute. Then $\sum_{k=1}^{\infty} a_{n_k} < \infty$ for any subsequence $(n_k)_{k\geq 1} \subseteq \mathbb{N}$. In particular, $\sum_{k=1}^{\infty} a_{\lfloor t_k \rfloor} < \infty$, which shows the result. To show the claim, let $m \in \mathbb{N}$ be such that $m\beta > p + 3$. Then the estimate $\exp(-x) \leq m! x^{-m}$ with $x \geq 0$ gives

$$n^{p+1} \exp(-\alpha n^{\beta}) \leq n^{p+1} \cdot m! \big(\alpha n^{\beta}\big)^{-m} = \frac{m!}{\alpha^m} n^{p+1-m\beta} \leq \frac{m!}{\alpha^m} n^{-2}. \tag{18}$$

Since $\sum_{n=1}^{\infty} n^{-2} < \infty$, indeed $\sum_{k=1}^{\infty} a_n < \infty$. □

Proposition A.1. Let f be a stationary Gaussian process and let w be a function that decays sufficiently quickly. Then

$$\lim_{t\to\infty} w(t)f(t) = 0 \tag{19}$$

almost surely.

Proof. For suppose not. Let A be measurable set with $\mathbb{P}(A) > 0$ on which

$$\limsup_{t\to\infty} |w(t)f(t)| = L > 0 \tag{20}$$

with $L = \infty$ allowed. Then there is a sequence $(t_n)_{n\geq 1} \subseteq \mathbb{R}$, $t_n \uparrow \infty$, such that

$$\lim_{n\to\infty} |w(t_n)f(t_n)| = L. \tag{21}$$

Since $t_n \uparrow \infty$, we may assume that $t_n \geq n$ for all n by passing to a subsequence. Set

$$B_n = \{|w(t_n)f(t_n)| \geq (C_0 + n)^{-1}\}, \tag{22}$$

where $C_0 > 0$ is chosen such that $C_0^{-1} < L$. By construction, $\mathbb{P}(B_n \text{ i.o.}) \geq \mathbb{P}(A) > 0$. We claim, however, that $\sum_{n=1}^{\infty} \mathbb{P}(B_n) < \infty$. Then, by the Borel–Cantelli lemma, $\mathbb{P}(B_n \text{ i.o.}) = 0$, which indeed is a contradiction. To show the claim, note that, by Markov's inequality,

$$\begin{aligned}
\mathbb{P}(B_n) &= \mathbb{P}(|w(t_n)f(t_n)| \geq (C_0 + n)^{-1}) && (23)\\
&\leq (C_0 + n)|w(t_n)|\mathbb{E}(|f(t_n)|) && (24)\\
&\leq C_1(C_0 + n)|w(t_n)| && (25)
\end{aligned}$$

where $C_1 = \mathbb{E}(|f(0)|) < \infty$. Thus $\sum_{n=1}^{\infty} \mathbb{P}(B_n) < \infty$, by Lem A.1. □

Lemma A.2. Let $f \geq 0$ and $g > 0$ be continuous. If

$$\lim_{t\to\infty} \frac{f(t)}{g(t)} = 0 \quad \text{and} \quad \lim_{t\to-\infty} \frac{f(t)}{g(t)} = 0, \tag{26}$$

then there exists a $C > 0$ such that Cg dominates f.

Proof. Let R be such that $|t| \geq R$ implies that $f(t)/g(t) < 1$. Then g dominates f on $(-\infty, -R] \cup [R, \infty)$. Since $[-R, R]$ is compact and f and g are continuous, on $[-R, R]$, f attains its maximum M and g its mimumum m, where $m > 0$ because $g > 0$. Setting $C = \max\{1, M/m\}$ then works. □

The filter h is generated according to

$$g \sim \mathcal{GP}(0, k_g), \qquad h(t) \,|\, g = w(t)g(t), \tag{27}$$

where

$$w(t) = \exp(-\alpha t^2), \qquad k_g(t - t') = \exp(-\gamma(t - t')^2). \tag{28}$$

Here g is stationary and, almost surely, has pathwise derivatives of all orders (*e.g.*, Theorem 4, Cambanis, 1973).

Let β be such that $0 < \beta < \alpha$. Then, almost surely,

$$\lim_{t\to\infty} \frac{|w(t)g(t)|}{\exp(-\beta t^2)} = 0 \quad \text{and} \quad \lim_{t\to-\infty} \frac{|w(t)g(t)|}{\exp(-\beta t^2)} = 0 \tag{29}$$

because $t \mapsto \exp(\beta t^2)w(t) = \exp(-(\alpha - \beta)t^2)$ decays sufficiently quickly. Thus, almost surely, there exists a $C > 0$ such that $t \mapsto C\exp(-\beta t^2)$ dominates wg, and $t \mapsto C\exp(-\beta t^2)$ is integrable and goes to zero at infinity.

Note that $t \mapsto \exp(\beta t^2)|w^{(n)}(t)|$, where $w^{(n)}$ is the n^{th} derivative of w, decays sufficiently quickly for all $n \in \mathbb{N}$. For any derivative of wg, use the product rule to expand and argue similarly to obtain a dominating function also of the form $t \mapsto C\exp(-\beta t^2)$.

In conclusion, almost surely, h and all derivatives of h are dominated by integrable functions that go to zero at infinity, and any product of these dominating functions is integrable. Therefore, almost surely, (1) h and all derivatives of h go to zero at infinity and, (2) any product of h and its derivatives is dominated by an integrable function.

B Proof of Proposition 2.2

Proof of Proposition 2.2. The proof proceeds like the proof for Prop 2.1. Let

$$z(r) = \int_0^\infty h(r+\tau)h(\tau)\,\mathrm{d}\tau. \tag{30}$$

Then

$$k_{f\,|\,h}(r) = z(|r|) = z(0) + z'(0)|r| + \tfrac{1}{2}z''(0)r^2 + O(|r|^3), \tag{31}$$

so sample paths of the CGPCM are almost surely everywhere differentiable if $z'(0) = 0$ and $z''(0) \neq 0$ and almost surely nowhere differentiable if $z'(0) \neq 0$ (Thms 3 and 4 from Cambanis, 1973). First, use integration by parts to find

$$z'(0) = \int_0^\infty h'(\tau)h(\tau)\,\mathrm{d}\tau = \cancel{\lim_{\tau\to\infty} h^2(\tau)} - h^2(0) - \int_0^\infty h(\tau)h'(\tau)\,\mathrm{d}\tau = -h^2(0) - z'(0) \tag{32}$$

almost surely. Thus $z'(0) = -\frac{1}{2}h^2(0)$ almost surely. Second, if $h(0) = 0$, again use integration by parts to find

$$z''(0) = \int_0^\infty h''(\tau)h(\tau)\,\mathrm{d}\tau = \cancel{\lim_{\tau\to\infty} h'(\tau)h(\tau)} - h'(0)h(0) - \int_0^\infty (h'(\tau))^2\,\mathrm{d}\tau = -\int_0^\infty (h'(\tau))^2\,\mathrm{d}\tau < 0. \tag{33}$$

almost surely. Therefore, if $h(0) = 0$, then $z'(0) = 0$ and $z''(0) \neq 0$ almost surely; and if $h(0) \neq 0$, then $z'(0) \neq 0$ almost surely. The result now follows. □

C The Causal Gaussian Process Convolution Model

C.1 Equivalent Formulations

Linear system formulation: Let k_h be the the following DEQ kernel:

$$k_h(t,t') = \tilde{\alpha}^2 e^{-\alpha t^2 - \alpha t'^2 - \gamma(t-t')^2}. \tag{34}$$

Then the linear system formulation of the CGPCM is given by the following generative model:

$$f(t)\,|\,h, x = \int_{-\infty}^t h(t-\tau)x(\tau)\,\mathrm{d}\tau \quad \text{where} \quad x \sim \mathcal{GP}(0, \delta(t-t')), \quad h \sim \mathcal{GP}(0, k_h(t,t')), \tag{35}$$

where $\delta(\,\cdot\,)$ denotes the Dirac delta function. Compared to the GPCM, the linear system formulation of the CGPCM uses a *causal* convolution operation.

Nonparametric kernel formulation: To derive the equivalent nonparametric kernel formulation, note that

$$\mathbb{E}[f(t)\,|\,h] = \int_{-\infty}^t h(t-\tau)\mathbb{E}[x(\tau)]\,\mathrm{d}\tau = 0. \tag{36}$$

Moreover,

$$\mathbb{E}[f(t)f(t')\,|\,h] = \int_{-\infty}^t \int_{-\infty}^{t'} h(t-\tau)h(t'-\tau')\mathbb{E}[x(\tau)x(\tau')]\,\mathrm{d}\tau'\,\mathrm{d}\tau = \int_{-\infty}^{t\wedge t'} h(t-\tau)h(t'-\tau)\,\mathrm{d}\tau. \tag{37}$$

If $t \leq t'$, then

$$\mathbb{E}[f(t)f(t')\,|\,h] = \int_{-\infty}^t h(t-\tau)h(t'-\tau)\,\mathrm{d}\tau = \int_0^\infty h(\tau)h(t'-t+\tau)\,\mathrm{d}\tau. \tag{38}$$

Similarly, if $t \geq t'$, then

$$\mathbb{E}[f(t)f(t')\,|\,h] = \int_{-\infty}^{t'} h(t-\tau)h(t'-\tau)\,\mathrm{d}\tau = \int_0^\infty h(t-t'+\tau)h(\tau)\,\mathrm{d}\tau. \tag{39}$$

Therefore, in any case,

$$\mathbb{E}[f(t)f(t')\,|\,h] = \int_0^\infty h(|t-t'|+\tau)h(\tau)\,\mathrm{d}\tau. \tag{40}$$

We conclude that the CGPCM is equivalent to the following generative model:

$$f\,|\,h \sim \mathcal{GP}\left(0, \int_0^\infty h(|t-t'|+\tau)h(\tau)\,\mathrm{d}\tau\right), \quad h \sim \mathcal{GP}(0, k_h(t,t')). \tag{41}$$

C.2 Local Behaviour

We argue that the irregularity of the sample paths of the CGPCM is controlled by the value of $|h(0)|$. Because h is smooth, for small $\varepsilon > 0$ and $\tau \in (0, \varepsilon)$, $h(\tau) \approx h(0)$, which means that

$$f(t+\varepsilon) - f(t) = \int_t^{t+\varepsilon} h(t-\tau)x(\tau)\,\mathrm{d}\tau \approx h(0)\int_t^{t+\varepsilon} x(\tau)\,\mathrm{d}\tau \overset{\mathrm{d}}{=} h(0)B_\varepsilon \tag{42}$$

where $\overset{\mathrm{d}}{=}$ denotes equality in distribution and $(B_t)_{t\geq 0}$ is a standard Brownian motion. Therefore, f can locally be approximated by a $|h(0)|$-scaled Brownian motion, which means that the magnitude of the irregular increments is controlled by the value of $|h(0)|$.

C.3 Inducing Points

For the filter h, define n_u inducing point inputs $\mathbf{t}_u$ initialised to uniformly spaced over $[-2\Delta, 2\tau_w]$ where Δ is the inter-point spacing and τ_w the extent of the filter (see Sec 2). Let the inducing points $\mathbf{u}$ then be

$$\mathbf{u} = (h(t_{u,1}), \ldots, h(t_{u,n_u})). \tag{43}$$

For the excitation signal x, we first define the inter-domain transformation

$$s(\tau) = \int_{-\infty}^{\infty} \tilde{\omega} e^{-\omega(\tau-t)^2} x(t)\,\mathrm{d}t. \tag{44}$$

Then define n_z inducing point inputs $\mathbf{t}_z$ initialised to uniformly spaced over a window containing the data and let the inducing points $\mathbf{z}$ be

$$\mathbf{z} = (s(t_{z,1}), \ldots, s(t_{z,n_u})). \tag{45}$$

With these choices, we have the covariances

$$k_{\mathbf{u}}(t) \coloneqq \mathbb{E}[h(t)h(\mathbf{t}_u)] = k_h(t, \mathbf{t}_u), \tag{46}$$

$$\mathbf{K}_{\mathbf{u}} \coloneqq \mathbb{E}[h(\mathbf{t}_u)h(\mathbf{t}_u)^\top] = k_h(\mathbf{t}_u, \mathbf{t}_u^\top), \tag{47}$$

$$k_{\mathbf{z}}(t) \coloneqq \mathbb{E}[x(t)s(\mathbf{t}_z)] = \int_{-\infty}^{\infty} \tilde{\omega} e^{-\omega(\mathbf{t}_z - t')^2} \mathbb{E}[x(t)x(t')]\,\mathrm{d}t' = \tilde{\omega} e^{-\omega(\mathbf{t}_z - t)^2}, \tag{48}$$

$$\mathbf{K}_{\mathbf{z}} \coloneqq \mathbb{E}[s(\mathbf{t}_z)s(\mathbf{t}_z)^\top] = \tilde{\omega}^2 \sqrt{\frac{\pi}{2\omega}} e^{-\frac{1}{2}\omega(\mathbf{t}_z - \mathbf{t}_z^\top)^2} \tag{49}$$

where the expression for $\mathbf{K}_{\mathbf{z}}$ follows from

$$\mathbb{E}[s(\mathbf{t}_z)s(\mathbf{t}_z)^\top] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \tilde{\omega}^2 e^{-\omega(\mathbf{t}_z - t)^2 - \omega(\mathbf{t}_z^\top - t')^2} \mathbb{E}[x(t)x(t')]\,\mathrm{d}t'\,\mathrm{d}t \tag{50}$$

$$= \int_{-\infty}^{\infty} \tilde{\omega}^2 e^{-\omega(\mathbf{t}_z - t)^2 - \omega(\mathbf{t}_z^\top - t)^2}\,\mathrm{d}t \tag{51}$$

$$= \tilde{\omega}^2 \sqrt{\frac{\pi}{2\omega}} e^{-\frac{1}{2}\omega(\mathbf{t}_z - \mathbf{t}_z^\top)^2}. \tag{52}$$

D The Rough Gaussian Process Convolution Model

D.1 Equivalent Formulations

Linear system formulation: Let k_h be the covariance function of white noise windowed by $w(t) = \tilde{\alpha} e^{-\alpha|t|}$:

$$k_h(t, t') = \tilde{\alpha}^2 e^{-\alpha|t| - \alpha|t'|} \delta(t - t') \tag{53}$$

where $\delta(\,\cdot\,)$ denotes the Dirac delta function. Let k_x be the covariance function of an Ornstein–Uhlenbeck process with length scale λ:

$$k_x(t, t') = e^{-\lambda|t-t'|}. \tag{54}$$

Then the linear system formulation of the RGPCM is given by the following generative model:

$$f(t) \,|\, h, x = \int_{-\infty}^{t} h(t-\tau)x(\tau)\,\mathrm{d}\tau \quad \text{where} \quad x \sim \mathcal{GP}(0, k_x(t,t')), \quad h \sim \mathcal{GP}(0, k_h(t,t')). \tag{55}$$

Compared to the (C)GPCM, the filter h has now a white noise prior and the input signal x is given by an OU process.

Nonparametric kernel formulation: To derive the equivalent nonparametric kernel formulation, note that

$$\mathbb{E}[f(t) \,|\, h] = \int_{-\infty}^{t} h(t-\tau)\mathbb{E}[x(\tau)]\,\mathrm{d}\tau = 0 \tag{56}$$

and

$$\mathbb{E}[f(t)f(t') \,|\, h] = \int_{-\infty}^{t}\int_{-\infty}^{t'} h(t-\tau)h(t'-\tau')\mathbb{E}[x(\tau)x(\tau')]\,\mathrm{d}\tau'\,\mathrm{d}\tau \tag{57}$$

$$= \int_{-\infty}^{t}\int_{-\infty}^{t'} h(t-\tau)h(t'-\tau')k_x(\tau-\tau')\,\mathrm{d}\tau'\,\mathrm{d}\tau \tag{58}$$

$$= \int_{0}^{\infty}\int_{0}^{\infty} h(\tau)h(\tau')k_x((t-t')-(\tau-\tau'))\,\mathrm{d}\tau'\,\mathrm{d}\tau. \tag{59}$$

We conclude that the RGPCM is equivalent to the following generative model:

$$f \,|\, h \sim \mathcal{GP}\bigg(0, \int_{0}^{\infty}\int_{0}^{\infty} h(\tau)h(\tau')k_x((t-t')-(\tau-\tau'))\,\mathrm{d}\tau'\,\mathrm{d}\tau\bigg), \quad h \sim \mathcal{GP}(0, k_h(t,t')). \tag{60}$$

Nonparametric spectral formulation: Associated with the the linear system (60) is the *frequency response* of the filter h

$$g(\omega) = \int_{0}^{\infty} e^{-i\omega t}h(t)\,\mathrm{d}t. \tag{61}$$

Note that g is a Gaussian process, since the Fourier transform in (61) is a linear operator. From the *spectral representation* of f, it then follows that the spectral density of f is given by

$$\phi_f(\omega) = |g(\omega)|^2\phi_x(\omega), \tag{62}$$

see Lindgren (2012). Effectively, the spectrum is a (random) modulation of the input signal's spectrum $\phi_x(\omega) = \frac{1}{\pi}\frac{\lambda}{\lambda^2+\omega^2}$. This can also be seen through the Fourier duality with the nonparameteric kernel

$$\mathbb{E}[f(t)f(t') \,|\, g] = \int e^{i\omega(t-t')}|g(\omega)|^2\phi_x(\omega)\,\mathrm{d}\omega. \tag{63}$$

Finally, we note that the spectral density of the fractional Ornstein–Uhlenbeck process with Hurst exponent H is given by

$$\phi_{\mathrm{fOU}}(\omega) = c|\omega|^{1-2H}\frac{\lambda}{\lambda^2+\omega^2} \tag{64}$$

where c is a normalising constant, see Cheridito et al. (2003). Hence, this is a parametric special case of the nonparameteric spectrum (62), namely $|g(\omega)|^2 = c\pi|\omega|^{1-2H}$.

D.2 Inducing Points

For the filter h, since it now enjoys a white noise prior, we first define an inter-domain transformation, which we choose the be *causal*:

$$s(\tau) = \int_{-\infty}^{\tau} \tilde{\gamma}e^{-\gamma|\tau-t|}h(t)\,\mathrm{d}t. \tag{65}$$

As we will see below, by choosing the inter-domain transform to be causal, the kernel of the inter-domain process s will take a simple form. Then define n_u inducing point inputs $\mathbf{t}_u$ initialised to uniformly spaced over $[0, \tau_w]$ where τ_w is the extent of the filter (see Sec 2) and let the inducing points $\mathbf{u}$ be

$$\mathbf{u} = (s(t_{u,1}), \ldots, s(t_{u,n_u})). \tag{66}$$

For the excitation signal x, we consider a collection of projections tailored for learning features in the spectral domain (Hensman et al., 2018). These features are defined on a window $[a, b]$ of interest. Typically, this window should contain the locations of all observed data points and also points where you want to make predictions. To define the features, let $M \in \mathbb{N}$ and consider the following basis functions:

$$\beta_m(t) = \begin{cases} 1 & \text{if } m = 0, \\ \cos(\omega_m(t-a)) & \text{if } 1 \leq m \leq M, \\ \sin(\omega_{m-M}(t-a)) & \text{if } M < m \leq 2M, \end{cases} \tag{67}$$

where the frequencies are harmonics on the interval $[a, b]$:

$$\omega_m = \frac{2\pi m}{b-a}, \quad m = 1, \ldots, M. \tag{68}$$

Denote the concatenation of all these features by $\boldsymbol{\beta}(t)$. We then let the inducing points $\mathbf{z}$ be

$$\mathbf{z} = (\langle x, \beta_0 \rangle_{\mathbb{H}}, \ldots, \langle x, \beta_{n_z} \rangle_{\mathbb{H}}) \tag{69}$$

where $n_z = 2M + 1$ and $\langle \cdot, \cdot \rangle_{\mathbb{H}}$ is the inner product corresponding to the reproducing kernel Hilbert space $\mathbb{H}$ associated with k_x. With these choices, we have the covariances

$$k_{\mathbf{u}}(t) \coloneqq \mathbb{E}[h(t)s(\mathbf{t}_u)] = \int_{-\infty}^{\mathbf{t}_u} \tilde{\gamma} e^{-\gamma|\mathbf{t}_u - t'|} \mathbb{E}[h(t)h(t')] \, \mathrm{d}t' = \tilde{\gamma} e^{-\gamma(t-\mathbf{t}_u)} \mathbb{1}(t \leq \mathbf{t}_u), \tag{70}$$

$$\mathbf{K}_{\mathbf{u}} \coloneqq \mathbb{E}[s(\mathbf{t}_u)s(\mathbf{t}_u)^\top] = \frac{\tilde{\gamma}^2}{2\gamma} e^{-\gamma|\mathbf{t}_u - \mathbf{t}_u^\top|}, \tag{71}$$

$$k_{\mathbf{z}}(t) \coloneqq \mathbb{E}[x(t)\langle x, \boldsymbol{\beta} \rangle_{\mathbb{H}}] = \langle \mathbb{E}[x(t)x], \boldsymbol{\beta} \rangle_{\mathbb{H}} = \langle k_x(t, \cdot), \boldsymbol{\beta} \rangle_{\mathbb{H}} = \boldsymbol{\beta}(t), \tag{72}$$

$$\mathbf{K}_{\mathbf{z}} \coloneqq \mathbb{E}[\langle x, \boldsymbol{\beta} \rangle_{\mathbb{H}} \langle x, \boldsymbol{\beta}^\top \rangle_{\mathbb{H}}] = \langle \langle \mathbb{E}[x(\cdot)x(\cdot)], \boldsymbol{\beta} \rangle_{\mathbb{H}}, \boldsymbol{\beta}^\top \rangle_{\mathbb{H}} = \langle \langle k_x(\cdot, \cdot), \boldsymbol{\beta} \rangle_{\mathbb{H}}, \boldsymbol{\beta}^\top \rangle_{\mathbb{H}} = \langle \boldsymbol{\beta}, \boldsymbol{\beta}^\top \rangle_{\mathbb{H}}, \tag{73}$$

where we use the reproducing property of k_x on $\mathbb{H}$. The expression for $\mathbf{K}_{\mathbf{u}}$ follows from

$$\mathbb{E}[s(\mathbf{t}_u)s(\mathbf{t}_u^\top)] = \int_{-\infty}^{\mathbf{t}_u} \int_{-\infty}^{\mathbf{t}_u^\top} \tilde{\gamma}^2 e^{-\gamma|\mathbf{t}_u - t| - \gamma|\mathbf{t}_u^\top - t'|} \mathbb{E}[h(t)h(t')] \, \mathrm{d}t' \, \mathrm{d}t \tag{74}$$

$$= \int_{-\infty}^{\mathbf{t}_u \wedge \mathbf{t}_u^\top} \tilde{\gamma}^2 e^{-\gamma(\mathbf{t}_u - t) - \gamma(\mathbf{t}_u^\top - t)} \, \mathrm{d}t \tag{75}$$

$$= \frac{\tilde{\gamma}^2}{2\gamma} e^{-\gamma(\mathbf{t}_u + \mathbf{t}_u^\top) - 2\gamma(\mathbf{t}_u \wedge \mathbf{t}_u^\top)} \tag{76}$$

$$= \frac{\tilde{\gamma}^2}{2\gamma} e^{-\gamma|\mathbf{t}_u - \mathbf{t}_u^\top|}. \tag{77}$$

Note that $\mathbf{K}_{\mathbf{z}}$ requires explicit computation of $\langle \beta_m, \beta_n \rangle_{\mathbb{H}}$ for all $m, n = 0, \ldots, M$, which can be done using an explicit expression for $\langle \cdot, \cdot \rangle_{\mathbb{H}}$; see Hensman et al. (2018) for details.

E Initialisation of the GPCM, CGPCM, and RGPCM

Let the subscript $\cdot_{\mathrm{ac}}$ refer to the GPCM, $\cdot_{\mathrm{c}}$ to the CGPCM, and $\cdot_{\mathrm{r}}$ to the RGPCM. In this section, we derive a comparable and fair initialisation for the three models. This initialisation follows from the following two requirements: (1) the prior marginal variance is unity and (2) the length scales of the prior mean covariance are equal.

The prior marginal variance of the models are as follows:

$$P_{\mathrm{c}} = \frac{1}{2}\tilde{\alpha}^2 \sqrt{\frac{\pi}{2\alpha}}, \quad P_{\mathrm{ac}} = \tilde{\alpha}^2 \sqrt{\frac{\pi}{2\alpha}}, \quad P_{\mathrm{r}} = \frac{\tilde{\alpha}^2}{2\alpha} \tag{78}$$

where we note that $P_\mathrm{c} = \frac{1}{2}P_\mathrm{ac}$: the causality constraint cuts the reach of the filter in half. Define $\tilde{\alpha}$ by requiring that the power is one. This gives

$$\tilde{\alpha}_\mathrm{c}^2 = 2\sqrt{\frac{2\alpha}{\pi}}, \quad \tilde{\alpha}_\mathrm{ac}^2 = \sqrt{\frac{2\alpha}{\pi}}, \quad \tilde{\alpha}_\mathrm{rc}^2 = 2\alpha. \tag{79}$$

With these choices for $\tilde{\alpha}$, we obtain the following prior mean covariance functions:

$$k_\mathrm{ac}(r) = \exp(-(\tfrac{1}{2}\alpha + \gamma)r^2), \tag{80}$$
$$k_\mathrm{c}(r) = (1 - \operatorname{erf}(\sqrt{\tfrac{1}{2}\alpha}|r|))\exp(-(\tfrac{1}{2}\alpha + \gamma)r^2), \tag{81}$$
$$k_\mathrm{r}(r) = \exp(-\lambda|r|) \tag{82}$$

where we note that k_c is a windowed version of k_ac. This window $r \mapsto 1 - \operatorname{erf}((\frac{1}{2}\alpha)^{\frac{1}{2}}|r|)$ is nondifferentiable at $r = 0$ and responsible for the nowhere differentiable sample paths of the CGPCM.

Define the length scale τ of a non-negative function $k\colon [0, \infty) \to \mathbb{R}$ by

$$\tau = \frac{1}{k(0)}\int_0^\infty k(r)\,\mathrm{d}r. \tag{83}$$

Then the length scales of the windows w_ac, w_c, w_r are given by

$$\tau_{w,\mathrm{ac}} = \sqrt{\frac{\pi}{4\alpha}}, \quad \tau_{w,\mathrm{c}} = \sqrt{\frac{\pi}{4\alpha}}, \quad \tau_{w,\mathrm{v}} = \frac{1}{\alpha} \tag{84}$$

and the length scales of the prior mean covariances are given by

$$\tau_{f,\mathrm{ac}} = \sqrt{\frac{\pi}{2(\alpha + 2\gamma)}}, \tag{85}$$
$$\tau_{f,\mathrm{c}} = \sqrt{\frac{2}{\pi(\alpha + 2\gamma)}}\tanh\left(\sqrt{\frac{\alpha + 2\gamma}{\alpha}}\right) \overset{\gamma \gg \alpha}{\approx} \sqrt{\frac{\pi}{2(\alpha + 2\gamma)}}, \tag{86}$$
$$\tau_{f,\mathrm{r}} = \frac{1}{\lambda} \tag{87}$$

where the approximation follows from that $\tanh(x) \approx \frac{1}{2}\pi$ for $x \gg 1$. If we fix τ_w and τ_f to given values for all models, we obtain

$$\alpha_\mathrm{ac} = \frac{\pi}{4}\frac{1}{\tau_w^2}, \quad \alpha_\mathrm{c} = \frac{\pi}{4}\frac{1}{\tau_w^2}, \quad \alpha_\mathrm{r} = \frac{1}{\tau_w}. \tag{88}$$

and

$$\gamma_\mathrm{ac,c} = \frac{\pi}{4}\frac{1}{\tau_f^2} - \frac{1}{2}\alpha_\mathrm{ac,c}, \quad \lambda_\mathrm{r} = \frac{1}{\tau_f}. \tag{89}$$

Intuitively, τ_f should be set to smallest length scale in the signal, and τ_w should be set to desired length of the filter, a bit larger than the largest length scale in the signal.

F Inference in the GPCM Family

F.1 Implementation

We provide a JAX (Bradbury et al., 2018) based Python implementation of the GPCM, CGPCM, and RGPCM at `github.com/wesselb/gpcm`. The package provides an `sklearn`-style `model.fit(t, y)`–`model.predict(t_new)` interface. The following is an example of fitting the RGPCM to data. Here, `window` refers to τ_w and `scale` refers to τ_f.

```
import numpy as np

from gpcm import GPCM, CGPCM, RGPCM
```

```
model = RGPCM(window=2, scale=1, noise=0.1, t=(0, 10))

# Sample from the prior.
t = np.linspace(0, 10, 100)
k, y = model.sample(t)

# Fit model to the sample.
model.fit(t, y)

# Compute the ELBO.
elbo = model.elbo(t, y)

# Make predictions.
posterior = model.condition(t, y)
mean, var = posterior.predict(t)
```

A key difficulty in the implementation is that the CGPCM requires evaluation and gradients of bivariate normal CDF; see App F.6.

F.2 Variational Approximation

In what follows, $\mathbf{u}$ are inducing points (or features) for the filter h and $\mathbf{z}$ are inducing points (or features) for the excitation signal x; see Tobar et al. (2015a) for the specification of $\mathbf{u}$ and $\mathbf{z}$ for the GPCM and Apps C.3 and D.2 for the specifications of $\mathbf{u}$ and $\mathbf{z}$ for respectively the CGPCM and RGPCM. We consider the variational approximation

$$q_\theta(h, x, \mathbf{u}, \mathbf{z}) = p_\theta(h \mid \mathbf{u}) p_\theta(x \mid \mathbf{z}) q(\mathbf{u}, \mathbf{z}) \tag{90}$$

where $q(\mathbf{z}, \mathbf{u})$ is a joint variational approximation over the inducing points. Given some data $\mathbf{y}$, to optimise $q(\mathbf{z}, \mathbf{u})$ and θ, we optimise the *evidence lower bound* (ELBO):

$$\mathcal{F}_\theta[q(\mathbf{u}, \mathbf{z})] = \mathbb{E}_q[\log p_\theta(\mathbf{y} \mid f)] - \mathrm{KL}[q(\mathbf{u}, \mathbf{z}) \,\|\, p_\theta(\mathbf{u}) p_\theta(\mathbf{z})]. \tag{91}$$

As the name suggests, the ELBO indeed forms a lower bound on the marginal likelihood, $\mathcal{F}_\theta[q(\mathbf{u}, \mathbf{z})] \leq \log p_\theta(\mathbf{y})$, and optimising $\mathcal{F}_\theta[q(\mathbf{u}, \mathbf{z})]$ corresponds to minimising the Kullback–Leibler divergence between our approximation of the posterior $q_\theta(h, x, \mathbf{u}, \mathbf{z})$ and the true posterior $p_\theta(h, x, \mathbf{u}, \mathbf{z} \mid \mathbf{y})$ (see, *e.g.*, Wainwright and Jordan, 2008). Key to the tractability of the ELBO is the observation that

$$\mathbb{E}_q[\log p_\theta(\mathbf{y} \mid f)] = \mathbb{E}_{q(\mathbf{u}, \mathbf{z})}[\mathbb{E}[\log p_\theta(\mathbf{y} \mid f) \mid \mathbf{u}, \mathbf{z}]] \tag{92}$$

where $\mathbb{E}[\log p_\theta(\mathbf{y} \mid f) \mid \mathbf{u}, \mathbf{z}]$ happens to be *conditionally quadratic* in $\mathbf{u}$ and $\mathbf{z}$; see App F.5.1. Throughout, we drop the dependency on θ and denote $\hat{\mathbf{u}} = \mathbf{K}_{\mathbf{u}}^{-1}\mathbf{u}$ and $\hat{\mathbf{z}} = \mathbf{K}_{\mathbf{z}}^{-1}\mathbf{z}$. In the remainder of this section, we consider two types of inference schemes: various mean-field schemes and a structured scheme.

F.3 Mean-Field Inference

In the mean-field inference scheme, the variational distribution is assumed to factorise

$$q(\mathbf{u}, \mathbf{z}) = q(\mathbf{u}) q(\mathbf{z}) \tag{93}$$

such that no dependency can occur between the variables. This leads to a simplifying factorisation of the ELBO:

$$\mathcal{F}_\theta[q(\mathbf{u})q(\mathbf{z})] = \mathbb{E}_q[\log p_\theta(\mathbf{y} \mid f)] - \mathrm{KL}[q(\mathbf{u}) \,\|\, p_\theta(\mathbf{u})] - \mathrm{KL}[q(\mathbf{z}) \,\|\, p_\theta(\mathbf{z})]. \tag{94}$$

Tobar et al. (2015a) assume Gaussians

$$q(\mathbf{u}) = \mathcal{N}(\mathbf{u}; \boldsymbol{\mu}_{\mathbf{u}}, \boldsymbol{\Sigma}_{\mathbf{u}}), \quad q(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_{\mathbf{z}}, \boldsymbol{\Sigma}_{\mathbf{z}}) \tag{95}$$

and optimise (94) with respect the means $(\boldsymbol{\mu}_{\mathbf{u}}, \boldsymbol{\mu}_{\mathbf{z}})$ and dense covariance matrices $(\boldsymbol{\Sigma}_{\mathbf{u}}, \boldsymbol{\Sigma}_{\mathbf{z}})$ using gradient-based optimisation. As Tobar et al. remark and App F.5.2 shows, the optimal mean-field approximations $q^*(\mathbf{u})$ and $q^*(\mathbf{z})$ are indeed of a Gaussian form.

Coordinate ascent: App F.5.2 not only shows that the optimal $q^*(\mathbf{u})$ and $q^*(\mathbf{z})$ are Gaussian, it also shows that, given $q(\mathbf{z})$ (resp. $q(\mathbf{u})$), the optimal $q^*(\mathbf{u})$ (resp. $q^*(\mathbf{z})$) can be computed explicitly. This gives rise to a coordinate ascent (CA) scheme which avoids gradient-based optimisation:

(1) Set $q_0(\mathbf{u}) = p(\mathbf{u})$.

(2) For $i = 1, \ldots, m$,

(2.a) set $q_i(\mathbf{z})$ to the optimal $q^*(\mathbf{z})$ given $q_{i-1}(\mathbf{u})$, and
(2.b) set $q_i(\mathbf{u})$ to the optimal $q^*(\mathbf{u})$ given $q_i(\mathbf{z})$.

The expression for the optimal $q^*(\mathbf{z})$ given $q(\mathbf{u})$ is computed in (140) and (141) and the expression for the optimal $q^*(\mathbf{u})$ given $q(\mathbf{z})$ is exactly analogous.

Collapsed mean-field: In practice, the coordinate ascent scheme converges much quicker than gradient-based optimisation of the mean-field ELBO. A downside of the coordinate ascent scheme, however, is that it only optimises the variational approximation: it cannot optimise the hyperparameters θ and inducing point inputs. For this, we propose to plug the optimal $q^*(\mathbf{z})$ given $q(\mathbf{u})$ back into the mean-field ELBO: $\mathcal{F}_\theta[q(\mathbf{u})q^*(\mathbf{u})]$. This *collapsed* mean-field ELBO does not depend on the variational distribution $q(\mathbf{z})$ anymore: it depends on many fewer variational parameters, which greatly accelerates gradient-based optimisation. The expression for the collapsed mean-field ELBO is computed in (143).

Computational complexity: The dominating computation of the three mean-field schemes is the computation of expressions of the form $\mathbf{I}_{\mathbf{uz}}(t_i)\mathbf{X}\mathbf{I}_{\mathbf{uz}}^\top(t_i)$ for all $i = 1, \ldots, n$ where $\mathbf{X}$ is some $n_z \times n_z$ matrix (*e.g.*, see (135)), which takes time $O(n(n_u n_z^2 + n_u^2 n_z))$. Note that the analogous expression $\mathbf{I}_{\mathbf{uz}}^\top(t_i)\mathbf{Y}\mathbf{I}_{\mathbf{uz}}(t_i)$ for some $n_u \times n_u$ matrix $\mathbf{Y}$ is equally expensive.

F.4 Structured Inference

In the structured inference scheme, there is no independence assumption nor a Gaussianity assumption on the variational distribution: we consider a general potentially non-Gaussian $q(\mathbf{u}, \mathbf{z})$ which can model arbitrary dependencies between $\mathbf{u}$ and $\mathbf{z}$. Let

$$g(\mathbf{u}, \mathbf{z}) = \exp \mathbb{E}[\log p(\mathbf{y} \mid f) \mid \mathbf{u}, \mathbf{z}], \tag{96}$$

$$Z^* = \int p(\mathbf{u})p(\mathbf{z})g(\mathbf{u}, \mathbf{z}) \, \mathrm{d}\mathbf{u} \, \mathrm{d}\mathbf{z}. \tag{97}$$

Then the ELBO can be written as

$$\mathcal{F}_\theta[q(\mathbf{u}, \mathbf{z})] = \log Z^* - \mathrm{KL}(q(\mathbf{u}, \mathbf{z}) \parallel \tfrac{1}{Z^*} p(\mathbf{u})p(\mathbf{z})g(\mathbf{u}, \mathbf{z})), \tag{98}$$

which means that the optimal $q^*(\mathbf{u}, \mathbf{z})$ is given by $q^*(\mathbf{u}, \mathbf{z}) = \frac{1}{Z^*} p(\mathbf{u})p(\mathbf{z})g(\mathbf{u}, \mathbf{z})$. As App F.5.1 explains, $(\mathbf{u}, \mathbf{z}) \mapsto \mathbb{E}[\log p(\mathbf{y} \mid f) \mid \mathbf{u}, \mathbf{z}]$ can be computed, but unfortunately is a *quartic* function. Therefore, although can we evaluate $g(\mathbf{u}, \mathbf{z})$, we cannot compute Z^* analytically, which means that we can only evaluate $q^*(\mathbf{u}, \mathbf{z})$ up to a normalising constant. Moreover, plugging $q^*(\mathbf{u}, \mathbf{z})$ back into the ELBO gives

$$\mathcal{F}_\theta[q^*(\mathbf{u}, \mathbf{z})] = \log Z^*, \tag{99}$$

which is intractable because Z^* cannot be computed. In the next paragraphs, we describe how these intractabilities can be navigated to enable inference and learning with the structured approximation.

Inference through Gibbs sampling $q^*(\mathbf{u}, \mathbf{z})$: The optimal $q^*(\mathbf{u}, \mathbf{z})$ can be factorised as follows:

$$q^*(\mathbf{u}, \mathbf{z}) = q^*(\mathbf{u})q^*(\mathbf{z} \mid \mathbf{u}) \tag{100}$$

where

$$q^*(\mathbf{u}) \propto p(\mathbf{u})Z(\mathbf{u}), \tag{101}$$

$$q^*(\mathbf{z} \mid \mathbf{u}) = \frac{1}{Z(\mathbf{u})} p(\mathbf{z})g(\mathbf{z}, \mathbf{u}), \tag{102}$$

$$Z(\mathbf{u}) = \int p(\mathbf{z})g(\mathbf{u}, \mathbf{z}) \, \mathrm{d}\mathbf{z}. \tag{103}$$

For $q^*(\mathbf{u})$, App F.5.4 shows that it can only be evaluated up to a normalising constant, unfortunately. However, crucially, App F.5.3 shows that $q^*(\mathbf{z} \,|\, \mathbf{u})$ takes the form of a Gaussian with mean and variance depending on $\mathbf{u}$, which means that, for a given $\mathbf{u}$, $q^*(\mathbf{z} \,|\, \mathbf{u})$ can be directly sampled from. Therefore, although we cannot evaluate $q^*(\mathbf{u}, \mathbf{z})$ nor directly sample from it, we can eventually generate samples from $q^*(\mathbf{u}, \mathbf{z})$ with the following Gibbs sampling scheme:

(1) Draw initial sample $\mathbf{u}^{(0)} \sim p(\mathbf{u})$.

(2) For $i = 1, \ldots, m$,

(2.a) sample $\mathbf{z}^{(i)} \sim q^*(\mathbf{z} \,|\, \mathbf{u}^{(i-1)})$, and
(2.b) sample $\mathbf{u}^{(i)} \sim q^*(\mathbf{u} \,|\, \mathbf{z}^{(i)})$.

With this sampling scheme, we are able to perform inference in the model whilst completely avoiding numerical variational optimisation. The expression for $q^*(\mathbf{z} \,|\, \mathbf{u})$ is computed in (149) and (150) and the expression for $q^*(\mathbf{u} \,|\, \mathbf{z})$ is exactly analogous.

Learning through approximation of $\frac{\mathrm{d}}{\mathrm{d}\theta}\mathcal{F}_\theta[q^*(\mathbf{u}, \mathbf{z})]$: If we substitute the factorisation (100) back into the ELBO, we get

$$\mathcal{F}_\theta[q^*(\mathbf{u}, \mathbf{z})] = \mathcal{F}_\theta[q^*(\mathbf{z} \,|\, \mathbf{u}) q^*(\mathbf{u})] = \mathbb{E}_{q^*(\mathbf{u})}[\log Z(\mathbf{u})] - \mathrm{KL}(q^*(\mathbf{u}) \,\|\, p(\mathbf{u})). \tag{104}$$

As before, $\mathcal{F}_\theta[q^*(\mathbf{z} \,|\, \mathbf{u}) q^*(\mathbf{u})]$ is intractable. However, it turns out that gradients of $\mathcal{F}_\theta[q^*(\mathbf{z} \,|\, \mathbf{u}) q^*(\mathbf{u})]$ with respect to θ and inducing points inputs can be computed:

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathcal{F}_\theta[q^*(\mathbf{u}, \mathbf{z})] = \frac{\partial}{\partial\theta} F_\theta[q^*(\mathbf{u}, \mathbf{z})] + \cancel{\left\langle \frac{\delta F_\theta}{\delta q(\mathbf{u}, \mathbf{z})}[q^*(\mathbf{u}, \mathbf{z})], \frac{\partial}{\partial\theta} q^*(\mathbf{u}, \mathbf{z}) \right\rangle} \tag{105}$$

where $\frac{\mathrm{d}}{\mathrm{d}\theta}$ denotes the total derivative with respect to θ, $\frac{\partial}{\partial\theta}$ the partial derivative with respect to θ, and $\frac{\delta}{\delta q(\mathbf{u}, \mathbf{z})}$ the functional derivative with respect to $q(\mathbf{u}, \mathbf{z})$. Here the second term cancels because $\frac{\delta F_\theta}{\delta q(\mathbf{u}, \mathbf{z})}[q^*(\mathbf{u}, \mathbf{z})] = 0$ by optimality of $q^*(\mathbf{u}, \mathbf{z})$. Therefore, using the above Gibbs sampling scheme to generate samples from $q^*(\mathbf{u})$, the following approximation is tractable:

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathcal{F}_\theta[q^*(\mathbf{u}, \mathbf{z})] \approx \frac{\partial}{\partial\theta} \frac{1}{m} \sum_{i=1}^{m} \log Z_\theta(\mathbf{u}^{(i)}) p_\theta(\mathbf{u}^{(i)}) \quad \text{where} \quad \mathbf{u}^{(i)} \overset{\text{i.i.d.}}{\sim} q^*(\mathbf{u}) \tag{106}$$

where we make the dependence of $Z_\theta(\mathbf{u})$ and $p_\theta(\mathbf{u})$ on θ explicit. By iterating the Gibbs sampling scheme and (106), we are able to perform stochastic gradient-based optimisation of θ and the inducing point inputs. The expression for $\log Z(\mathbf{u}) p(\mathbf{u})$ is computed in (156).

Evidence approximation through a lower bound on $\mathcal{F}_\theta[q^*(\mathbf{u}, \mathbf{z})]$: Although we have described how gradients of $\mathcal{F}_\theta[q^*(\mathbf{u}, \mathbf{z})]$ can be approximated, some applications may require an approximation of the value of $\mathcal{F}_\theta[q^*(\mathbf{u}, \mathbf{z})]$. To tractably approximate $\mathcal{F}_\theta[q^*(\mathbf{u}, \mathbf{z})]$, we propose the following lower bound:

$$\mathcal{F}_\theta[q^*(\mathbf{u}, \mathbf{z})] \geq \mathcal{F}_\theta[q^*_{\mathrm{MF}}(\mathbf{u}) q^*(\mathbf{z} \,|\, \mathbf{u})] = \mathbb{E}_{q^*_{\mathrm{MF}}(\mathbf{u})}[\log Z(\mathbf{u})] - \mathrm{KL}(q^*_{\mathrm{MF}}(\mathbf{u}) \,\|\, p(\mathbf{u})) \tag{107}$$

where $q^*_{\mathrm{MF}}(\mathbf{u})$ is the optimal mean-field approximation of $q(\mathbf{u})$ obtained with the coordinate ascent procedure. The expectation can be approximated with a Monte Carlo approximation by sampling from $q^*_{\mathrm{MF}}(\mathbf{u})$. The expression for $\mathcal{F}_\theta[q(\mathbf{u}) q^*(\mathbf{z} \,|\, \mathbf{u})]$ for an arbitrary $q(\mathbf{u})$ is computed in (155).

Computational complexity: Like the mean-field schemes, the structured inference scheme also takes $O(n(n_u n_z^2 + n_u^2 n_z))$ time. However, for the Gibbs sampling scheme, by paying an upfront cost of $O(n(n_u n_z^2 + n_u^2 n_z))$ once, every Gibbs sample iteration can be performed in $O(n n_u n_z)$ time, which is a dramatic speed-up over $O(n(n_u n_z^2 + n_u^2 n_z))$. This speed-up makes it possible to always run the Gibbs sampler until convergence without excessive computational expense.

F.5 Computations

In this section, we give detailed derivations of all remaining computations.

F.5.1 Conditional Expectation of the Likelihood $\mathbb{E}[\log p(\mathbf{y} \mid f) \mid \mathbf{u}, \mathbf{z}]$

To begin with, compute

$$\mathbb{E}[f(t) \mid \mathbf{u}, \mathbf{z}] = \int_{-\infty}^{t} \mathbb{E}[h(t-\tau) \mid \mathbf{u}]\mathbb{E}[x(\tau) \mid \mathbf{z}] \, \mathrm{d}\tau = \hat{\mathbf{u}}^\top \int_{-\infty}^{t} k_{\mathbf{u}}(t-\tau) k_{\mathbf{z}}^\top(\tau) \, \mathrm{d}\tau \, \hat{\mathbf{z}} = \hat{\mathbf{u}}^\top \mathbf{I}_{\mathbf{uz}}(t) \hat{\mathbf{z}} \tag{108}$$

where we define

$$\mathbf{I}_{\mathbf{uz}}(t) \coloneqq \int_{-\infty}^{t} k_{\mathbf{u}}(t-\tau) k_{\mathbf{z}}^\top(\tau) \, \mathrm{d}\tau. \tag{109}$$

Write

$$\mathbb{E}[h(t)h(t') \mid \mathbf{u}] = k_h(t,t') - k_{\mathbf{u}}^\top(t) \mathbf{K}_{\mathbf{u}}^{-1} k_{\mathbf{u}}(t') + (k_{\mathbf{u}}^\top(t)\hat{\mathbf{u}})^2 = k_h(t,t') + k_{\mathbf{u}}^\top(t) \mathbf{M}_{\mathbf{u}} k_{\mathbf{u}}(t') \tag{110}$$

where $\mathbf{M}_{\mathbf{u}} = \hat{\mathbf{u}}\hat{\mathbf{u}}^\top - \mathbf{K}_{\mathbf{u}}^{-1}$. Then

$$\begin{aligned}
\mathbb{E}[f^2(t) \mid \mathbf{u}, \mathbf{z}] &= \int_{-\infty}^{t} \int_{-\infty}^{t'} \mathbb{E}[h(t-\tau)h(t-\tau') \mid \mathbf{u}]\mathbb{E}[x(\tau)x(\tau') \mid \mathbf{z}] \, \mathrm{d}\tau' \, \mathrm{d}\tau && (111)\\
&= \int_{-\infty}^{t} \int_{-\infty}^{t'} (k_h(t-\tau, t-\tau') + k_{\mathbf{u}}^\top(t-\tau) \mathbf{M}_{\mathbf{u}} k_{\mathbf{u}}(t-\tau'))(k_x(\tau,\tau') + k_{\mathbf{z}}^\top(\tau) \mathbf{M}_{\mathbf{z}} k_{\mathbf{z}}(\tau')) \, \mathrm{d}\tau' \, \mathrm{d}\tau && (112)\\
&= T_1(t) + T_2(t) + T_3(t) + T_4(t) && (113)
\end{aligned}$$

where

$$\begin{aligned}
T_1(t) &\coloneqq \int_{-\infty}^{t} \int_{-\infty}^{t} k_h(t-\tau, t-\tau') k_x(\tau,\tau') \, \mathrm{d}\tau' \, \mathrm{d}\tau && (114)\\
T_2(t) &\coloneqq \int_{-\infty}^{t} \int_{-\infty}^{t} k_h(t-\tau, t-\tau') k_{\mathbf{z}}^\top(\tau) \mathbf{M}_{\mathbf{z}} k_{\mathbf{z}}(\tau') \, \mathrm{d}\tau' \, \mathrm{d}\tau && (115)\\
&= \operatorname{tr} \mathbf{M}_{\mathbf{z}} \int_{-\infty}^{t} \int_{-\infty}^{t} k_h(t-\tau, t-\tau') k_{\mathbf{z}}(\tau') k_{\mathbf{z}}^\top(\tau) \, \mathrm{d}\tau' \, \mathrm{d}\tau && (116)\\
T_3(t) &\coloneqq \int_{-\infty}^{t} \int_{-\infty}^{t} k_{\mathbf{u}}^\top(t-\tau) \mathbf{M}_{\mathbf{u}} k_{\mathbf{u}}(t-\tau') k_x(\tau,\tau') \, \mathrm{d}\tau' \, \mathrm{d}\tau && (117)\\
&= \operatorname{tr} \mathbf{M}_{\mathbf{u}} \int_{-\infty}^{t} \int_{-\infty}^{t} k_{\mathbf{u}}(t-\tau') k_{\mathbf{u}}^\top(t-\tau) k_x(\tau,\tau') \, \mathrm{d}\tau' \, \mathrm{d}\tau && (118)\\
T_4(t) &\coloneqq \int_{-\infty}^{t} \int_{-\infty}^{t} k_{\mathbf{u}}^\top(t-\tau) \mathbf{M}_{\mathbf{u}} k_{\mathbf{u}}(t-\tau') k_{\mathbf{z}}^\top(\tau) \mathbf{M}_{\mathbf{z}} k_{\mathbf{z}}(\tau') \, \mathrm{d}\tau' \, \mathrm{d}\tau && (119)\\
&= \operatorname{tr} \mathbf{M}_{\mathbf{z}} \int_{-\infty}^{t} k_{\mathbf{z}}(\tau) k_{\mathbf{u}}^\top(t-\tau) \, \mathrm{d}\tau \, \mathbf{M}_{\mathbf{u}} \int_{-\infty}^{t} k_{\mathbf{u}}(t-\tau') k_{\mathbf{z}}^\top(\tau') \, \mathrm{d}\tau' && (120)\\
&= \operatorname{tr} \mathbf{M}_{\mathbf{z}} \mathbf{I}_{\mathbf{uz}}^\top(t) \mathbf{M}_{\mathbf{u}} \mathbf{I}_{\mathbf{uz}}(t). && (121)
\end{aligned}$$

Further define

$$\begin{aligned}
I_{hx}(t) &\coloneqq \int_{-\infty}^{t} \int_{-\infty}^{t} k_h(t-\tau, t-\tau') k_x(\tau,\tau') \, \mathrm{d}\tau' \, \mathrm{d}\tau, && (122)\\
\mathbf{I}_{h\mathbf{z}}(t) &\coloneqq \int_{-\infty}^{t} \int_{-\infty}^{t} k_h(t-\tau, t-\tau') k_{\mathbf{z}}(\tau') k_{\mathbf{z}}^\top(\tau) \, \mathrm{d}\tau' \, \mathrm{d}\tau, && (123)\\
\mathbf{I}_{\mathbf{u}x}(t) &\coloneqq \int_{-\infty}^{t} \int_{-\infty}^{t} k_{\mathbf{u}}(t-\tau') k_{\mathbf{u}}^\top(t-\tau) k_x(\tau,\tau') \, \mathrm{d}\tau' \, \mathrm{d}\tau. && (124)
\end{aligned}$$

Then

$$T_1(t) = I_{hx}(t), \tag{125}$$
$$T_2(t) = \langle \mathbf{M}_\mathbf{z}, \mathbf{I}_{h\mathbf{z}}(t) \rangle \tag{126}$$
$$= \hat{\mathbf{z}}^\top \mathbf{I}_{h\mathbf{z}}(t) \hat{\mathbf{z}} - \langle \mathbf{K}_\mathbf{z}^{-1}, \mathbf{I}_{h\mathbf{z}}(t) \rangle, \tag{127}$$
$$T_3(t) = \langle \mathbf{M}_\mathbf{u}, \mathbf{I}_{\mathbf{u}x}(t) \rangle \tag{128}$$
$$= \hat{\mathbf{u}}^\top \mathbf{I}_{\mathbf{u}x}(t) \hat{\mathbf{u}} - \langle \mathbf{K}_\mathbf{u}^{-1}, \mathbf{I}_{\mathbf{u}x}(t) \rangle, \tag{129}$$
$$T_4(t) = \langle \mathbf{M}_\mathbf{u}, \mathbf{I}_{\mathbf{uz}}(t) \mathbf{M}_\mathbf{z} \mathbf{I}_{\mathbf{uz}}^\top(t) \rangle, \tag{130}$$
$$= \langle \hat{\mathbf{u}}\hat{\mathbf{u}}^\top - \mathbf{K}_\mathbf{u}^{-1}, \mathbf{I}_{\mathbf{uz}}(t) (\hat{\mathbf{z}}\hat{\mathbf{z}}^\top - \mathbf{K}_\mathbf{z}^{-1}) \mathbf{I}_{\mathbf{uz}}^\top(t) \rangle, \tag{131}$$
$$= (\hat{\mathbf{u}}^\top \mathbf{I}_{\mathbf{uz}}(t) \hat{\mathbf{z}})^2 - \hat{\mathbf{u}}^\top \mathbf{I}_{\mathbf{uz}}(t) \mathbf{K}_\mathbf{z}^{-1} \mathbf{I}_{\mathbf{uz}}^\top(t) \hat{\mathbf{u}} \\ - \hat{\mathbf{z}}^\top \mathbf{I}_{\mathbf{uz}}^\top(t) \mathbf{K}_\mathbf{u}^{-1} \mathbf{I}_{\mathbf{uz}}(t) \hat{\mathbf{z}} + \langle \mathbf{K}_\mathbf{u}^{-1}, \mathbf{I}_{\mathbf{uz}}(t) \mathbf{K}_\mathbf{z}^{-1} \mathbf{I}_{\mathbf{uz}}^\top(t) \rangle. \tag{132}$$

Aggregate

$$\mathbf{A}(t) \coloneqq \mathbf{I}_{\mathbf{u}x}(t) - \mathbf{I}_{\mathbf{uz}}(t) \mathbf{K}_\mathbf{z}^{-1} \mathbf{I}_{\mathbf{uz}}^\top(t), \tag{133}$$
$$\mathbf{B}(t) \coloneqq \mathbf{I}_{h\mathbf{z}}(t) - \mathbf{I}_{\mathbf{uz}}^\top(t) \mathbf{K}_\mathbf{u}^{-1} \mathbf{I}_{\mathbf{uz}}(t), \tag{134}$$
$$c(t) \coloneqq I_{hx}(t) - \langle \mathbf{K}_\mathbf{u}^{-1}, \mathbf{I}_{\mathbf{u}x}(t) \rangle - \langle \mathbf{K}_\mathbf{z}^{-1}, \mathbf{I}_{h\mathbf{z}}(t) \rangle + \langle \mathbf{K}_\mathbf{u}^{-1}, \mathbf{I}_{\mathbf{uz}}(t) \mathbf{K}_\mathbf{z}^{-1} \mathbf{I}_{\mathbf{uz}}^\top(t) \rangle. \tag{135}$$

Then

$$\mathbb{E}[f^2(t) \mid \mathbf{u}, \mathbf{z}] = \hat{\mathbf{u}}^\top \mathbf{A}(t) \hat{\mathbf{u}} + \hat{\mathbf{z}}^\top \mathbf{B}(t) \hat{\mathbf{z}} + (\hat{\mathbf{u}}^\top \mathbf{I}_{\mathbf{uz}}(t) \hat{\mathbf{z}})^2 + c(t). \tag{136}$$

With $\mathbb{E}[f(t) \mid \mathbf{u}, \mathbf{z}]$ and $\mathbb{E}[f^2(t) \mid \mathbf{u}, \mathbf{z}]$ computed, we can compute $\mathbb{E}[\log p(\mathbf{y} \mid f) \mid \mathbf{u}, \mathbf{z}]$:

$$\mathbb{E}[\log p(\mathbf{y} \mid f) \mid \mathbf{u}, \mathbf{z}] = \sum_{i=1}^n \left[-\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \big(\mathbb{E}[f^2(t_i) \mid \mathbf{u}, \mathbf{z}] - 2 y_i \mathbb{E}[f(t_i) \mid \mathbf{u}, \mathbf{z}] + y_i^2 \big) \right] \tag{137}$$
$$= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|\mathbf{y}\|_2^2 \\ - \frac{1}{2\sigma^2} \sum_{i=1}^n \big[\hat{\mathbf{u}}^\top \mathbf{A}(t_i) \hat{\mathbf{u}} + \hat{\mathbf{z}}^\top \mathbf{B}(t_i) \hat{\mathbf{z}} + (\hat{\mathbf{u}}^\top \mathbf{I}_{\mathbf{uz}}(t_i) \hat{\mathbf{z}})^2 + c(t_i) - 2 y_i \hat{\mathbf{u}}^\top \mathbf{I}_{\mathbf{uz}}(t_i) \hat{\mathbf{z}} \big]. \tag{138}$$
$$= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \left[\|\mathbf{y}\|_2^2 + \hat{\mathbf{u}}^\top \sum_{i=1}^n \mathbf{A}(t_i) \hat{\mathbf{u}} + \sum_{i=1}^n c(t_i) \right] \\ - \frac{1}{2\sigma^2} \sum_{i=1}^n \big[\hat{\mathbf{z}}^\top \big(\mathbf{B}(t_i) + \mathbf{I}_{\mathbf{uz}}^\top(t_i) \hat{\mathbf{u}}\hat{\mathbf{u}}^\top \mathbf{I}_{\mathbf{uz}}(t_i) \big) \hat{\mathbf{z}} - 2 y_i \hat{\mathbf{u}}^\top \mathbf{I}_{\mathbf{uz}}(t_i) \hat{\mathbf{z}} \big]. \tag{139}$$

Crucially, observe that both $\mathbf{u} \mapsto \mathbb{E}[\log p(\mathbf{y} \mid f) \mid \mathbf{u}, \mathbf{z}]$ and $\mathbf{z} \mapsto \mathbb{E}[\log p(\mathbf{y} \mid f) \mid \mathbf{u}, \mathbf{z}]$ are quadratic functions. We therefore say that $\mathbb{E}[\log p(\mathbf{y} \mid f) \mid \mathbf{u}, \mathbf{z}]$ is *conditionally quadratic* in $\mathbf{u}$ and $\mathbf{z}$. The function $(\mathbf{u}, \mathbf{z}) \mapsto \mathbb{E}[\log p(\mathbf{y} \mid f) \mid \mathbf{u}, \mathbf{z}]$, however, is quartic rather than quadratic. It remains to compute the integrals $I_{hx}(t)$, $\mathbf{I}_{h\mathbf{z}}(t)$, $\mathbf{I}_{\mathbf{u}x}(t)$, and $\mathbf{I}_{\mathbf{uz}}(t)$. We will do this for the GPCM and CGPCM in App F.6 and for the RGPCM in App F.7.

F.5.2 Optimal Mean-Field $q^*(\mathbf{z})$ Given $q(\mathbf{u})$ and the Collapsed Mean-Field ELBO

We borrow the result from the next section, which computes the optimal $q^*(\hat{\mathbf{z}} \mid \hat{\mathbf{u}})$ and the partial structured ELBO. To instead compute the optimal $q^*(\mathbf{z})$ given $q(\mathbf{u})$ in the mean-field scheme and the collapsed mean-field ELBO, simply take an additional expectation over $q(\mathbf{u})$. In particular, let $q(\hat{\mathbf{u}}) = \mathcal{N}(\boldsymbol{\mu}_{\hat{\mathbf{u}}}, \boldsymbol{\Sigma}_{\hat{\mathbf{u}}})$. Then $q^*(\hat{\mathbf{z}}) = \mathcal{N}(\boldsymbol{\mu}_{\hat{\mathbf{z}}}, \boldsymbol{\Sigma}_{\hat{\mathbf{z}}})$ where

$$\boldsymbol{\Sigma}_{\hat{\mathbf{z}}}^{-1} = \mathbf{K}_\mathbf{z} + \frac{1}{\sigma^2} \left[\sum_{i=1}^n \mathbf{B}(t_i) + \sum_{i=1}^n \mathbf{I}_{\mathbf{uz}}^\top(t_i) (\boldsymbol{\Sigma}_{\hat{\mathbf{u}}} + \boldsymbol{\mu}_{\hat{\mathbf{u}}} \boldsymbol{\mu}_{\hat{\mathbf{u}}}^\top) \mathbf{I}_{\mathbf{uz}}(t_i) \right], \tag{140}$$
$$\boldsymbol{\Sigma}_{\hat{\mathbf{z}}}^{-1} \boldsymbol{\mu}_{\hat{\mathbf{z}}} = \frac{1}{\sigma^2} \sum_{i=1}^n y_i \mathbf{I}_{\mathbf{uz}}^\top(t_i) \boldsymbol{\mu}_{\hat{\mathbf{u}}}. \tag{141}$$

Similarly,

$$\log \int p(\mathbf{z}) \exp \mathbb{E}_{q(\mathbf{u})}[\mathbb{E}[\log p(\mathbf{y} \mid f) \mid \mathbf{u}, \mathbf{z}]] \, \mathrm{d}\mathbf{z}$$
$$= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \left[\|\mathbf{y}\|_2^2 + \hat{\mathbf{u}}^\top \sum_{i=1}^n \mathbf{A}(t_i) \hat{\mathbf{u}} + \sum_{i=1}^n c(t_i) \right] + \frac{1}{2} \log |\mathbf{\Sigma}_{\hat{\mathbf{z}}}| + \frac{1}{2} \log |\mathbf{K}_{\mathbf{z}}| + \frac{1}{2} \boldsymbol{\mu}_{\hat{\mathbf{z}}}^\top \mathbf{\Sigma}_{\hat{\mathbf{z}}}^{-1} \boldsymbol{\mu}_{\hat{\mathbf{z}}}, \quad (142)$$

so

$$\mathcal{F}_\theta[q^*(\mathbf{z}) q(\mathbf{u})]$$
$$= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \left[\|\mathbf{y}\|_2^2 + \operatorname{tr} \left[(\mathbf{\Sigma}_{\hat{\mathbf{u}}} + \boldsymbol{\mu}_{\hat{\mathbf{u}}} \boldsymbol{\mu}_{\hat{\mathbf{u}}}^\top) \sum_{i=1}^n \mathbf{A}(t_i) \right] + \sum_{i=1}^n c(t_i) \right]$$
$$+ \frac{1}{2} \log |\mathbf{\Sigma}_{\hat{\mathbf{z}}}| + \frac{1}{2} \log |\mathbf{K}_{\mathbf{z}}| + \frac{1}{2} \boldsymbol{\mu}_{\hat{\mathbf{z}}}^\top \mathbf{\Sigma}_{\hat{\mathbf{z}}}^{-1} \boldsymbol{\mu}_{\hat{\mathbf{z}}} - \mathrm{KL}(q(\mathbf{u}) \,\|\, p(\mathbf{u})). \quad (143)$$

F.5.3 Optimal Structured $q^*(\mathbf{z} \mid \mathbf{u})$ and the Partially Structured ELBO

To begin with, expand

$$\log p(\mathbf{z}) + \mathbb{E}[\log p(\mathbf{y} \mid f) \mid \mathbf{u}, \mathbf{z}] \quad (144)$$
$$= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2} \log|2\pi \mathbf{K}_{\mathbf{z}}| - \frac{1}{2} \hat{\mathbf{z}}^\top \mathbf{K}_{\mathbf{z}} \hat{\mathbf{z}} - \frac{1}{2\sigma^2} \left[\|\mathbf{y}\|_2^2 + \hat{\mathbf{u}}^\top \sum_{i=1}^n \mathbf{A}(t_i) \hat{\mathbf{u}} + \sum_{i=1}^n c(t_i) \right]$$
$$- \frac{1}{2\sigma^2} \sum_{i=1}^n \left[\hat{\mathbf{z}}^\top \left(\mathbf{B}(t_i) + \mathbf{I}_{\mathbf{uz}}^\top(t_i) \hat{\mathbf{u}} \hat{\mathbf{u}}^\top \mathbf{I}_{\mathbf{uz}}(t_i) \right) \hat{\mathbf{z}} - 2 y_i \hat{\mathbf{u}}^\top \mathbf{I}_{\mathbf{uz}}(t_i) \hat{\mathbf{z}} \right]. \quad (145)$$

Note that

$$\log \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma}) = -\frac{1}{2} \log|2\pi\mathbf{\Sigma}| - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \quad (146)$$
$$= -\frac{1}{2} \log|2\pi\mathbf{\Sigma}| - \frac{1}{2} (\mathbf{x}^\top \mathbf{\Sigma}^{-1} \mathbf{x} - 2\boldsymbol{\mu}^\top \mathbf{\Sigma}^{-1} \mathbf{x}) - \frac{1}{2} \boldsymbol{\mu}^\top \mathbf{\Sigma}^{-1} \boldsymbol{\mu}, \quad (147)$$

so

$$-\frac{1}{2} (\mathbf{x}^\top \mathbf{\Sigma}^{-1} \mathbf{x} - 2\boldsymbol{\mu}^\top \mathbf{\Sigma}^{-1} \mathbf{x}) = \log \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma}) + \frac{1}{2} \log|2\pi\mathbf{\Sigma}| + \frac{1}{2} \boldsymbol{\mu}^\top \mathbf{\Sigma}^{-1} \boldsymbol{\mu}. \quad (148)$$

Hence, observe that $q^*(\hat{\mathbf{z}} \mid \hat{\mathbf{u}}) = \mathcal{N}(\boldsymbol{\mu}_{\hat{\mathbf{z}}}, \mathbf{\Sigma}_{\hat{\mathbf{z}}}^{-1})$ where

$$\mathbf{\Sigma}_{\hat{\mathbf{z}}}^{-1} = \mathbf{K}_{\mathbf{z}} + \frac{1}{\sigma^2} \left[\sum_{i=1}^n \mathbf{B}(t_i) + \sum_{i=1}^n \mathbf{I}_{\mathbf{uz}}^\top(t_i) \hat{\mathbf{u}} \hat{\mathbf{u}}^\top \mathbf{I}_{\mathbf{uz}}(t_i) \right], \quad (149)$$

$$\mathbf{\Sigma}_{\hat{\mathbf{z}}}^{-1} \boldsymbol{\mu}_{\hat{\mathbf{z}}} = \frac{1}{\sigma^2} \sum_{i=1}^n y_i \mathbf{I}_{\mathbf{uz}}^\top(t_i) \hat{\mathbf{u}}. \quad (150)$$

Thus

$$\log p(\mathbf{z}) + \mathbb{E}[\log p(\mathbf{y} \mid f) \mid \mathbf{u}, \mathbf{z}]$$
$$= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2} \log|2\pi \mathbf{K}_{\mathbf{z}}| - \frac{1}{2\sigma^2} \left[\|\mathbf{y}\|_2^2 + \hat{\mathbf{u}}^\top \sum_{i=1}^n \mathbf{A}(t_i) \hat{\mathbf{u}} + \sum_{i=1}^n c(t_i) \right]$$
$$+ \log \mathcal{N}(\hat{\mathbf{z}}; \boldsymbol{\mu}_{\hat{\mathbf{z}}}, \mathbf{\Sigma}_{\hat{\mathbf{z}}}) + \frac{1}{2} \log |2\pi\mathbf{\Sigma}_{\hat{\mathbf{z}}}| + \frac{1}{2} \boldsymbol{\mu}_{\hat{\mathbf{z}}}^\top \mathbf{\Sigma}_{\hat{\mathbf{z}}}^{-1} \boldsymbol{\mu}_{\hat{\mathbf{z}}} \quad (151)$$
$$= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \left[\|\mathbf{y}\|_2^2 + \hat{\mathbf{u}}^\top \sum_{i=1}^n \mathbf{A}(t_i) \hat{\mathbf{u}} + \sum_{i=1}^n c(t_i) \right]$$
$$+ \log \mathcal{N}(\hat{\mathbf{z}}; \boldsymbol{\mu}_{\hat{\mathbf{z}}}, \mathbf{\Sigma}_{\hat{\mathbf{z}}}) + \frac{1}{2} \log |\mathbf{\Sigma}_{\hat{\mathbf{z}}}| - \frac{1}{2} \log |\mathbf{K}_{\mathbf{z}}| + \frac{1}{2} \boldsymbol{\mu}_{\hat{\mathbf{z}}}^\top \mathbf{\Sigma}_{\hat{\mathbf{z}}}^{-1} \boldsymbol{\mu}_{\hat{\mathbf{z}}}. \quad (152)$$

We now apply exp, integrate over $\mathrm{d}\mathbf{z} = |\mathbf{K}_\mathbf{z}|\,\mathrm{d}\hat{\mathbf{z}}$, and apply log to find

$$\begin{aligned}
&\log \int p(\mathbf{z}) \exp \mathbb{E}[\log p(\mathbf{y} \mid f) \mid \mathbf{u}, \mathbf{z}] \,\mathrm{d}\mathbf{z} \\
&\quad = \log Z(\mathbf{u}) && (153)\\
&\quad = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left[\|\mathbf{y}\|_2^2 + \hat{\mathbf{u}}^\top \sum_{i=1}^n \mathbf{A}(t_i)\hat{\mathbf{u}} + \sum_{i=1}^n c(t_i)\right] + \frac{1}{2}\log|\mathbf{\Sigma}_{\hat{\mathbf{z}}}| + \frac{1}{2}\log|\mathbf{K}_\mathbf{z}| + \frac{1}{2}\boldsymbol{\mu}_{\hat{\mathbf{z}}}^\top \mathbf{\Sigma}_{\hat{\mathbf{z}}}^{-1}\boldsymbol{\mu}_{\hat{\mathbf{z}}}, && (154)
\end{aligned}$$

so

$$\begin{aligned}
&\mathcal{F}_\theta[q(\mathbf{u})q^*(\mathbf{u} \mid \mathbf{z})] \\
&\quad = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left[\|\mathbf{y}\|_2^2 + \sum_{i=1}^n c(t_i)\right] - \frac{1}{2}\log|\mathbf{K}_\mathbf{z}| \\
&\qquad + \mathbb{E}_{q(\mathbf{u})}\left[-\frac{1}{2\sigma^2}\hat{\mathbf{u}}^\top \sum_{i=1}^n \mathbf{A}(t_i)\hat{\mathbf{u}} + \frac{1}{2}\log|\mathbf{\Sigma}_{\hat{\mathbf{z}}}(\hat{\mathbf{u}})| + \frac{1}{2}\boldsymbol{\mu}_{\hat{\mathbf{z}}}^\top(\hat{\mathbf{u}})\mathbf{\Sigma}_{\hat{\mathbf{z}}}^{-1}(\hat{\mathbf{u}})\boldsymbol{\mu}_{\hat{\mathbf{z}}}(\hat{\mathbf{u}})\right] - \mathrm{KL}(q(\mathbf{u}), p(\mathbf{u})). && (155)
\end{aligned}$$

Here the expectation can be approximated using Monte Carlo.

F.5.4 Optimal Structured $q^*(\mathbf{u})$ and Gradients for the Structured ELBO

Expand

$$\begin{aligned}
&\log q^*(\mathbf{u}) \simeq \log p(\mathbf{u})Z(\mathbf{u}) \\
&\qquad = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left[\|\mathbf{y}\|_2^2 + \hat{\mathbf{u}}^\top \sum_{i=1}^n \mathbf{A}(t_i)\hat{\mathbf{u}} + \sum_{i=1}^n c(t_i)\right] \\
&\qquad\quad - \frac{1}{2}\log|2\pi\mathbf{K}_\mathbf{u}| - \frac{1}{2}\hat{\mathbf{u}}^\top\mathbf{K}_\mathbf{u}\hat{\mathbf{u}} + \frac{1}{2}\log|\mathbf{\Sigma}_{\hat{\mathbf{z}}}(\hat{\mathbf{u}})| - \frac{1}{2}\log|\mathbf{K}_\mathbf{z}| + \frac{1}{2}\boldsymbol{\mu}_{\hat{\mathbf{z}}}^\top(\hat{\mathbf{u}})\mathbf{\Sigma}_{\hat{\mathbf{z}}}^{-1}(\hat{\mathbf{u}})\boldsymbol{\mu}_{\hat{\mathbf{z}}}(\hat{\mathbf{u}}) && (156)
\end{aligned}$$

where make the dependence of $\boldsymbol{\mu}_{\hat{\mathbf{z}}}(\hat{\mathbf{u}})$ and $\mathbf{\Sigma}_{\hat{\mathbf{z}}}(\hat{\mathbf{u}})$ on $\hat{\mathbf{u}}$ explicit. We see that $q^*(\hat{\mathbf{u}})$ can be evaluated up to a normalising constant.

F.6 Integrals for the GPCM and CGPCM

Rather than explicitly computing the integrals $I_{hx}(t)$, $\mathbf{I}_{h\mathbf{z}}(t)$, $\mathbf{I}_{\mathbf{u}x}(t)$, and $\mathbf{I}_{\mathbf{uz}}(t)$, we note that, for the GPCM and CGPCM, all these constitute integrals of exponentiated quadratic forms. We therefore implement a small computer algebra system (CAS) which is able to symbolically solve the integrals and implement the solutions in JAX. (For the GPCM, the integrals $I_{hx}(t)$, $\mathbf{I}_{h\mathbf{z}}(t)$, $\mathbf{I}_{\mathbf{u}x}(t)$, and $\mathbf{I}_{\mathbf{uz}}(t)$ are defined with different limits, but the CAS is able to handle that.) For the CGPCM, this requires availability of the bivariate normal CDF. For this, we use the FORTRAN implementation `TVPACK` by Alan Genz[8], parallelise the implementation in C++ using OpenMP, and hook the result into JAX's JIT compiler with manually defined gradients.

Below is an example of using the CAS to compute the integral $\mathbf{I}_{h\mathbf{z}}(t)$:

```
import numpy as np

from gpcm.exppoly import ExpPoly, const, var

t = np.linspace(0, 10, 100)
t_z = np.linspace(0, 10, 10)

alpha = 1
alpha_t = 1
gamma = 2
omega = 1
omega_t = 1
```

[8] `http://www.math.wsu.edu/faculty/genz/software/software.html`

```
def k_h(t1, t2):
    return alpha_t ** 2 * ExpPoly(
        -(const(alpha) * (t1 ** 2 + t2 ** 2) + const(gamma) * (t1 - t2) ** 2),
    )


def k_xs(t1, t2):
    return omega_t * ExpPoly(-const(omega) * (t1 - t2) ** 2)


expq = (
    k_h(var("t") - var("tau1"), var("t") - var("tau2"))
    * k_xs(var("tau1"), var("t_z_1"))
    * k_xs(var("t_z_2"), var("tau2"))
)

I_hz = expq.integrate_box(
    ("tau1", -np.inf, var("t")),
    ("tau2", -np.inf, var("t")),
    t=t[:, None, None],
    t_z_1=t_z[None, :, None],
    t_z_2=t_z[None, None, :],
)
```

F.7 Integrals for the RGPCM

In what follows, recall that

$$k_h(t,t') = w(t)w(t')k_g(t,t') \quad \text{with} \quad w(t) = \tilde{\alpha}e^{-\alpha|t|}, \quad k_g(t,t') = \delta(t-t') \tag{157}$$

and

$$k_x(t,t') = e^{-\lambda|t-t'|}. \tag{158}$$

F.7.1 Integral $I_{hx}(t)$

Compute

$$I_{hx}(t,t') = \int_{-\infty}^{t}\int_{-\infty}^{t'} k_h(t-\tau,t'-\tau')k_x(\tau,\tau')\,\mathrm{d}\tau'\,\mathrm{d}\tau \tag{159}$$

$$= \int_{-\infty}^{t}\int_{-\infty}^{t'} w(t-\tau)w(t'-\tau')k_g(t-\tau,t'-\tau')k_x(\tau,\tau')\,\mathrm{d}\tau'\,\mathrm{d}\tau \tag{160}$$

$$= \int_{0}^{\infty} w^2(\tau)k_x(t-\tau,t'-\tau)\,\mathrm{d}\tau \tag{161}$$

$$= \frac{\tilde{\alpha}^2}{2\alpha}e^{-\lambda|t-t'|}, \tag{162}$$

so

$$I_{hx} = I_{hx}(t,t) = \frac{\tilde{\alpha}^2}{2\alpha}. \tag{163}$$

F.7.2 Integral $\mathbf{I}_{h\mathbf{z}}(t)$

Denote the $(m,n)^{\text{th}}$ element of $\mathbf{I}_{h\mathbf{z}}(t)$ by

$$[I_{h\mathbf{z}}(t)]_{m,n} = \int_{-\infty}^{t}\int_{-\infty}^{t} w(t-\tau)w(t-\tau')k_g(t-\tau,t-\tau')k_{z_m}(\tau)k_{z_n}(\tau')\,\mathrm{d}\tau'\,\mathrm{d}\tau \tag{164}$$

$$= \int_{-\infty}^{t} w^2(t-\tau)k_{z_m}(\tau)k_{z_n}(\tau')\,\mathrm{d}\tau \tag{165}$$

$$=: I_{m,n}(t). \tag{166}$$

For $m = n = 0$, we have

$$I_{0,0}(t) = \int_a^t w^2(t-\tau)\,\mathrm{d}\tau + \int_{-\infty}^a w^2(t-\tau)e^{-2\lambda(a-\tau)}\,\mathrm{d}\tau \tag{167}$$

$$= \frac{\tilde{\alpha}^2}{2\alpha}(1 - e^{-2\alpha(t-a)}) + \frac{\tilde{\alpha}^2}{2(\alpha+\lambda)}e^{-2\alpha(t-a)} \tag{168}$$

$$= \frac{\tilde{\alpha}^2}{2\alpha} - \frac{\lambda\tilde{\alpha}^2}{2\alpha(\alpha+\lambda)}e^{-2\alpha(t-a)}. \tag{169}$$

For $m = 0$ and $1 \leq n \leq M$, we have cosine features:

$$I_{0,n:\cos}(t) = \int_a^t w^2(t-\tau)\cos(\omega_n(\tau-a))\,\mathrm{d}\tau + \int_{-\infty}^a w^2(t-\tau)e^{-2\lambda(a-\tau)}\,\mathrm{d}\tau \tag{170}$$

$$= \frac{\tilde{\alpha}^2}{4\alpha^2+\omega_n^2}\left[2\alpha\left(\cos(\omega_n(t-a)) - e^{-2\alpha(t-a)}\right) + \omega_n\sin(\omega_n(t-a))\right] + \frac{\tilde{\alpha}^2}{2(\alpha+\lambda)}e^{-2\alpha(t-a)} \tag{171}$$

$$= I_{0,0}(t) \quad \text{if} \quad \omega_n = 0. \tag{172}$$

Similarly, for $m, n \leq M$,

$$I_{m:\cos,n:\cos}(t) = \int_a^t w^2(t-\tau)\cos(\omega_m(\tau-a))\cos(\omega_n(\tau-a))\,\mathrm{d}\tau + \int_{-\infty}^a w^2(t-\tau)e^{-2\lambda(a-\tau)}\,\mathrm{d}\tau \tag{173}$$

$$= \frac{1}{2}\int_a^t w^2(t-\tau)\cos(\omega_{m-n}(\tau-a))\,\mathrm{d}\tau + \frac{1}{2}\int_a^t w^2(t-\tau)\cos(\omega_{m+n}(\tau-a))\,\mathrm{d}\tau + \int_{-\infty}^a w^2(t-\tau)e^{-2\lambda(a-\tau)}\,\mathrm{d}\tau \tag{174}$$

$$= \frac{1}{2}\left(I_{0,(n-m):\cos}(t) + I_{0,(n+m):\cos}(t)\right) \tag{175}$$

where we use that $\omega_m \pm \omega_n = \frac{2\pi}{b-a}(m \pm n) = \omega_{m\pm n}$. In the following, for $m > M$, recall that we adjust the frequency according to the construction of the variational Fourier features (see (67)):

$$k_{z_m}(\tau) = \beta_m(t) = \sin(\omega_{m-M}(\tau-a)). \tag{176}$$

For $m = 0$ and $n > M$, we have sines:

$$I_{0,n:\sin}(t) = \int_a^t w^2(t-\tau)\sin(\omega_{n-M}(\tau-a))\,\mathrm{d}\tau \tag{177}$$

$$= \int_0^{\omega_{n-M}(t-a)} w^2(t-a-\tau/\omega_{n-M})\sin(\tau)\frac{1}{\omega_{n-M}}\,\mathrm{d}\tau \tag{178}$$

$$= \frac{\tilde{\alpha}^2}{\omega_{n-M}}\int_0^{\omega_{n-M}(t-a)} e^{-2\frac{\alpha}{\omega_{n-M}}(\omega_{n-M}(t-a)-\tau)}\sin(\tau)\,\mathrm{d}\tau \tag{179}$$

$$= \frac{\tilde{\alpha}^2}{4\alpha^2+\omega_{n-M}^2}\left[\omega_{n-M}(e^{-2\alpha(t-a)} - \cos(\omega_{n-M}(t-a))) + 2\alpha\sin(\omega_{n-M}(t-a))\right]. \tag{180}$$

Next, for $m, n > M$,

$$I_{m:\sin,n:\sin}(t) = \int_a^t w^2(t-\tau)\sin(\omega_{m-M}(\tau-a))\sin(\omega_{n-M}(\tau-a))\,\mathrm{d}\tau \tag{181}$$

$$= \frac{1}{2}\int_a^t w^2(t-\tau)\cos(\omega_{m-n}(\tau-a))\,\mathrm{d}\tau - \frac{1}{2}\int_a^t w^2(t-\tau)\cos(\omega_{m+n-2M}(\tau-a))\,\mathrm{d}\tau \tag{182}$$

$$= \frac{1}{2}\left(I_{0,(n-m):\cos} - I_{0,(n+m-2M):\cos}\right). \tag{183}$$

Finally, for $0 < m \leq M$ and $n > M$, we have both cosines and sines:

$$I_{m:\cos,n:\sin}(t) = \int_a^t w^2(t-\tau)\cos(\omega_m(\tau-a))\sin(\omega_{n-M}(\tau-a))\,\mathrm{d}\tau \tag{184}$$

$$= \frac{1}{2}\int_a^t w^2(t-\tau)\sin(\omega_{m+(n-M)}(\tau-a))\,\mathrm{d}\tau + \frac{1}{2}\int_a^t w^2(t-\tau)\sin(\omega_{(n-M)-m}(\tau-a))\,\mathrm{d}\tau \tag{185}$$

$$= \frac{1}{2}\left(I_{0,(n+m):\sin} + I_{0,(n-m):\sin}\right). \tag{186}$$

F.7.3 Integral $\mathbf{I}_{\mathbf{u}x}(t)$

Denote the $(m,n)^{\text{th}}$ element of $\mathbf{I}_{\mathbf{u}x}(t,t')$ by

$$[\mathbf{I}_{\mathbf{u}x}(t,t')]_{m,n} = \int_{-\infty}^t \int_{-\infty}^{t'} w(t-\tau)w(t'-\tau')k_{u_m}(t-\tau)k_{u_n}(t'-\tau')k_x(\tau,\tau')\,\mathrm{d}\tau'\,\mathrm{d}\tau \tag{187}$$

$$= \int_0^\infty \int_0^\infty w(\tau)w(\tau')k_{u_m}(\tau)k_{u_n}(\tau')k_x(t-\tau,t'-\tau')\,\mathrm{d}\tau'\,\mathrm{d}\tau \tag{188}$$

$$=: I_{m,n}(t,t'). \tag{189}$$

Simplify

$$I_{m,n}(t,t') = \int_0^{t_{u,n}} \int_0^{t_{u,m}} \tilde{\alpha}^2\tilde{\gamma}^2 e^{-\alpha(\tau+\tau')-\gamma(t_{u,m}-\tau)-\gamma(t_{u,n}-\tau')-\lambda|(\tau-\tau')-(t-t')|}\,\mathrm{d}\tau\,\mathrm{d}\tau' \tag{190}$$

$$= \tilde{\alpha}^2\tilde{\gamma}^2 e^{-\gamma(t_{u,m}+t_{u,n})}\int_0^{t_{u,n}} \int_0^{t_{u,m}} e^{(\gamma-\alpha)(\tau+\tau')-\lambda|(\tau-\tau')-(t-t')|}\,\mathrm{d}\tau\,\mathrm{d}\tau'. \tag{191}$$

Note that $I_{m,n}(t,t)$ is invariant of t. The integral $I_{m,n}(t,t)$ can be computed with the following propositions.

Proposition F.1. Suppose that $a \geq 0$ and $b \geq 0$. Then

$$\int_0^a \int_0^b e^{c(\tau+\tau')-d|\tau-\tau'|}\,\mathrm{d}\tau\,\mathrm{d}\tau' = \frac{1}{c^2-d^2}\left(1 + \frac{d}{c}\left(1-e^{2c(a\wedge b)}\right) - e^{ca-d|a|} - e^{cb-d|b|} + e^{c(a+b)-d|a-b|}\right). \tag{192}$$

Proof. Suppose that $b \geq a$. Then $a - b = -|a - b|$ and $a = a \wedge b$. We simply calculate:

$$\int_0^a \int_0^b e^{c(\tau+\tau')-d|\tau-\tau'|} \,\mathrm{d}\tau \,\mathrm{d}\tau' = \int_0^a \int_0^{\tau'} e^{c(\tau+\tau')-d(\tau'-\tau)} \,\mathrm{d}\tau \,\mathrm{d}\tau' + \int_0^a \int_{\tau'}^b e^{c(\tau+\tau')-d(\tau-\tau')} \,\mathrm{d}\tau \,\mathrm{d}\tau' \tag{193}$$

$$= \int_0^a e^{(c-d)\tau'} \int_0^{\tau'} e^{(c+d)\tau} \,\mathrm{d}\tau \,\mathrm{d}\tau' + \int_0^a e^{(c+d)\tau'} \int_{\tau'}^b e^{(c-d)\tau} \,\mathrm{d}\tau \,\mathrm{d}\tau' \tag{194}$$

$$= \frac{1}{c+d} \int_0^a e^{(c-d)\tau'} \left(e^{(c+d)\tau'} - 1 \right) \mathrm{d}\tau' + \frac{1}{c-d} \int_0^a e^{(c+d)\tau'} \left(e^{(c-d)b} - e^{(c-d)\tau'} \right) \mathrm{d}\tau' \tag{195}$$

$$= \frac{1}{c+d} \int_0^a \left(e^{2c\tau'} - e^{(c-d)\tau'} \right) \mathrm{d}\tau' + \frac{1}{c-d} \int_0^a \left(e^{(c-d)b} e^{(c+d)\tau'} - e^{2c\tau'} \right) \mathrm{d}\tau' \tag{196}$$

$$= \left[\frac{1}{2c(c+d)} \left(e^{2ca} - 1 \right) - \frac{1}{c^2 - d^2} \left(e^{(c-d)a} - 1 \right) \right] + \left[\frac{e^{(c-d)b}}{c^2 - d^2} \left(e^{(c+d)a} - 1 \right) - \frac{1}{2c(c-d)} \left(e^{2ca} - 1 \right) \right] \tag{197}$$

$$= \frac{1}{2c} \left(\frac{1}{c+d} - \frac{1}{c-d} \right) \left(e^{2ca} - 1 \right) + \frac{1}{c^2 - d^2} \left(1 - e^{(c-d)a} - e^{(c-d)b} + e^{(c+d)a+(c-d)b} \right) \tag{198}$$

$$= \frac{d}{c} \frac{1}{c^2 - d^2} \left(1 - e^{2ca} \right) + \frac{1}{c^2 - d^2} \left(1 - e^{ca-da} - e^{cb-db} + e^{c(a+b)+d(a-b)} \right). \tag{199}$$

□

Proposition F.2. Suppose that $ab \leq 0$. Then

$$\int_0^a \int_0^b e^{c(\tau+\tau')-d|\tau-\tau'|} \,\mathrm{d}\tau \,\mathrm{d}\tau' = \frac{1}{c^2 - d^2} \left(1 - e^{ca-|d|a} - e^{cb-|d|b} + e^{c(a+b)-d|a-b|} \right). \tag{200}$$

We can finally use the symmetry in c to get the result for all $a, b \in \mathbb{R}$:

Proposition F.3. For all $a, b \in \mathbb{R}$,

$$\int_0^a \int_0^b e^{c(\tau+\tau')-d|\tau-\tau'|} \,\mathrm{d}\tau \,\mathrm{d}\tau' = \frac{1}{c^2 - d^2} \left(\mathbb{1}(ab \geq 0) \frac{d \operatorname{sign}(a)}{c} \left(1 - e^{2c \operatorname{sign}(a)(|a| \wedge |b|)} \right) + 1 - e^{ca-d|a|} - e^{cb-d|b|} + e^{c(a+b)-d|a-b|} \right). \tag{201}$$

Putting everything together, we have the result

$$I_{m,n}(t,t) = \frac{\tilde{\alpha}^2 \tilde{\gamma}^2 e^{-\gamma(t_{u,m}+t_{u,n})}}{(\gamma-\alpha)^2 - \lambda^2} \left(\frac{\lambda}{\gamma - \alpha} \left(1 - e^{2(\gamma-\alpha)(t_{u,m} \wedge t_{u,n})} \right) + 1 - e^{(\gamma-\alpha-\lambda)t_{u,m}} - e^{(\gamma-\alpha-\lambda)t_{u,n}} + e^{(\gamma-\alpha)(t_{u,m}+t_{u,n})-\lambda|t_{u,m}-t_{u,n}|} \right).$$

F.7.4 Integral $\mathbf{I}_{\mathbf{uz}}(t)$

Denote the $(m,k)^{\text{th}}$ element of $\mathbf{I}_{\mathbf{uz}}(t)$ by

$$[\mathbf{I}_{\mathbf{uz}}(t)]_{m,k} = \int_0^\infty w(\tau) k_{u_m}(\tau) k_{z_k}(t-\tau) \,\mathrm{d}\tau =: I_{m,k}(t). \tag{202}$$

Simplify

$$I_{m,k}(t) = \tilde{\alpha}\tilde{\gamma} \int_0^{t_{u,m}} e^{-\alpha\tau-\gamma(t_{u,m}-\tau)} k_{z_k}(t-\tau)\,\mathrm{d}\tau \tag{203}$$

$$= \tilde{\alpha}\tilde{\gamma} e^{-\gamma t_{u,m}} \int_0^{t_{u,m}} e^{(\gamma-\alpha)\tau} k_{z_k}(t-\tau)\,\mathrm{d}\tau \tag{204}$$

$$= \tilde{\alpha}\tilde{\gamma} e^{-\gamma t_{u,m}+(\gamma-\alpha)t} \int_{t-t_{u,m}}^{t} e^{(-\gamma+\alpha)\tau} k_{z_k}(\tau)\,\mathrm{d}\tau. \tag{205}$$

We analyse the result case by case. Denote

$$I(l,u,k) = \int_l^u e^{(-\gamma+\alpha)\tau} k_{z_k}(\tau)\,\mathrm{d}\tau. \tag{206}$$

The case $k=0$, $a \le l \le b$, and $a \le u \le b$:

$$I(l,u,0) = \int_l^u e^{(-\gamma+\alpha)\tau}\,\mathrm{d}\tau = \frac{1}{-\gamma+\alpha}\left(e^{(-\gamma+\alpha)u} - e^{(-\gamma+\alpha)l}\right). \tag{207}$$

The case $0 \le k \le K$ and $l,u \le a$:[9]

$$I(l,u,k) = \int_l^u e^{(-\gamma+\alpha)\tau-\lambda(a-\tau)}\,\mathrm{d}\tau = \frac{e^{-\lambda a}}{-\gamma+\alpha+\lambda}\left(e^{(-\gamma+\alpha+\lambda)u} - e^{(-\gamma+\alpha+\lambda)l}\right). \tag{208}$$

The case $0 \le k \le K$ and $l,u \ge b$:[9]

$$I(l,u,k) = \int_l^u e^{(-\gamma+\alpha)\tau-\lambda(\tau-b)}\,\mathrm{d}\tau = \frac{e^{\lambda b}}{-\gamma+\alpha-\lambda}\left(e^{(-\gamma+\alpha-\lambda)u} - e^{(-\gamma+\alpha-\lambda)l}\right). \tag{209}$$

The case $M < k \le 2K$ and either $l,u \le a$ or $l,u \ge b$:[9]

$$I(l,u,k) = 0. \tag{210}$$

The case $1 \le k \le M$, $a \le l \le b$, and $a \le u \le b$:

$$I(l,u,k) + iI(l,u,k+M) = \int_l^u e^{(-\gamma+\alpha)\tau+i\omega_k(\tau-a)}\,\mathrm{d}\tau \tag{211}$$

$$= \frac{e^{-i\omega_k a}}{-\gamma+\alpha+i\omega_k}\left(e^{(-\gamma+\alpha+i\omega_k)u} - e^{(-\gamma+\alpha+i\omega_k)l}\right) \tag{212}$$

$$= \frac{-\gamma+\alpha-i\omega_k}{(-\gamma+\alpha)^2+\omega_k^2}\left(e^{(-\gamma+\alpha)u+i\omega_k(u-a)} - e^{(-\gamma+\alpha)l+i\omega_k(l-a)}\right), \tag{213}$$

which shows that

$$\begin{aligned}((-\gamma+\alpha)^2+\omega_k^2)I(l,u,k) = {} & e^{(-\gamma+\alpha)u}[(-\gamma+\alpha)\cos(\omega_k(u-a)) + \omega_k \sin(\omega_k(u-a))] \\ & - e^{(-\gamma+\alpha)l}[(-\gamma+\alpha)\cos(\omega_k(l-a)) + \omega_k \sin(\omega_k(l-a))]\end{aligned} \tag{214}$$

and

$$\begin{aligned}((-\gamma+\alpha)^2+\omega_k^2)I(l,u,k+M) = {} & e^{(-\gamma+\alpha)u}[(\gamma-\alpha)\sin(\omega_k(u-a)) + \omega_k \cos(\omega_k(u-a))] \\ & - e^{(-\gamma+\alpha)l}[(\gamma-\alpha)\sin(\omega_k(l-a)) + \omega_k \cos(\omega_k(l-a))].\end{aligned} \tag{215}$$

[9] We could define $\mathbf{t}_u$ and $[a,b]$ such that this case never happens.