Function-space Distributions over Kernels

Greg Benton [1] **Jayson Salkey** [1] **Wesley Maddox** [1] **Julio Albinati** [2] **Andrew Gordon Wilson** [1]

Abstract

Gaussian processes are flexible function approximators, with inductive biases controlled by a covariance kernel. Learning the kernel is the key to representation learning and strong predictive performance. In this paper, we develop *functional kernel learning* (FKL) to directly infer functional posteriors over kernels. In particular, we place a transformed Gaussian process over a spectral density, to induce a non-parametric distribution over kernel functions. The resulting approach enables learning of rich representations, with support for any stationary kernel, uncertainty over the values of the kernel, and an interpretable specification of a prior directly over kernels, without requiring sophisticated initialization or manual intervention. We perform inference through elliptical slice sampling, which is especially well suited to marginalizing posteriors with the strongly correlated priors typical to function space modelling. We develop our approach for non-uniform, large-scale, multi-task, and multidimensional data, and show promising performance in a wide range of settings, including interpolation, extrapolation, and kernel recovery experiments.

1. Introduction & Background

Gaussian Processes (GPs) are highly flexible non-parametric models that, with simple assumptions, are capable of detailed pattern discovery. Accurate exact inference can proceed with an assumed parametric form on the covariance structure (i.e. RBF or periodic kernels); in some cases the limitations of these easily parameterized covariance functions are too severe for highly accurate prediction, or they require careful composition that cannot be deployed to general tasks (Williams & Rasmussen, 2006).

In many cases it is sufficient to assume the process is weakly stationary as this still admits many of the most popular choices for parametric form of the kernel (RBF, Matèrn, periodic, etc). In this setting we can simplify the kernel $k(x, x')$ to just a function of the distance between these points: $k(\tau) = k(||x - x‘||)$.

Recent work has utilized the spectral decomposition of kernels - their Fourier transforms - to generate prior distributions that provide support to broad classes of covariance functions (Wilson & Adams, 2013; Remes et al., 2017). While existing methods are effective at uncovering correlation structure in data, they are still reliant on parametric forms. Alternatively, we approach kernel learning from a non-parametric perspective, modelling the spectral density of the covariance function with a latent Gaussian process.

We propose this method as a drop-in replacement for covariance priors like the spectral mixture kernel by placing a GP prior over the spectral representation of covariance structure. This imbues the process of modeling the covariance with the same benefits seen in GP modeling of data (flexibility, simplicity, and a nonparametric approach). In the end we are able to define one model over the covariance of a GP that is able to realize and sample from a wide array of outcomes, and provides not only uncertainty in the estimation of data but in the estimation of the underlying kernel.

2. Methods

2.1. Spectral Transformations of Kernel Functions

Bochner's Theorem Bochner (1959) describes positive definite functions, $k(\tau)$, as the Fourier transform of finite Lebesgue measure on the real line. Thus, $k(\tau)$ is the covariance of a stationary integrable process on $\mathbb{R}$ if and only if

$$k(\tau) = \int_{\mathbb{R}} e^{2\pi i \omega \tau} S(\omega) d\omega, \tag{1}$$

for a positive, finite density $S(\omega)$. This transformation is simply the Fourier dual of the spectral density, $S(s)$ and implies invertibility. If $S(\omega)$ is known, $k(\tau)$ can be computed.

Uniqueness is guaranteed by the Wiener-Khintchine Theorem (Eq. 4.6 of Williams & Rasmussen (2006)). Note that $S(\omega)$ is an un-normalized Lebesgue measure, and so the pointwise variance (or outputscale) of the process is given by $k(0) = \int_{\mathbb{R}} d\mu(\omega)$.

[1]Cornell University [2]Microsoft. Correspondence to: Greg Benton <gwb67@cornell.edu>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

In the context of stationary covariance Gaussian processes, $k(\tau)$ is real-valued and can be made symmetric without loss of generality. Equation 1 simplifies to

$$k(\tau) = \int_{[0,\infty)} \cos(2\pi\tau\omega)S(\omega)d\omega, \tag{2}$$

by utilizing the complex exponential and oddness of sine. This is a slight simplification of Equations 4.7 and 4.8 in Williams & Rasmussen (2006).

For finitely sampled data, we can only identify frequencies up to π/Δ, where Δ is the largest spacing between neighboring points. This further restricts the integral to be over the space, $[0, \pi/\Delta]$, giving

$$k(\tau) = \int_{[0,\pi/\Delta]} \cos(2\pi\tau\omega)S(\omega)d\omega. \tag{3}$$

For an arbitrary density $S(\omega)$ this does not admit an analytic form of $k(\tau)$, however simple numerical integration schemes allow $k(\tau)$ to be approximated with high accuracy. We will utilize the trapezoid rule for this work, giving the approximate form of this integral as,

$$k(\tau) \approx \frac{\Delta_\omega}{2} \sum_{i=1}^{I} \cos(2\pi\tau\omega_i)S(\omega_i) + \cos(2\pi\tau\omega_{i-1})S(\omega_{i-1}), \tag{4}$$

assuming the spectrum is sampled at W evenly spaced points ω_i that are Δ_ω units apart in the frequency domain. This is a safe assumption as the choice of where to sample the spectrum is purely a modeling choice and can be fixed at the onset of any experimentation.

2.2. Specification of Latent Density

Since the transformation in Equation 1 is unique, we propose to fit a Gaussian process to the log-spectral density of a kernel $k(\tau)$. The log transformation assures that the spectral representation is non-negative and corresponds to a positive definite kernel.

Functional Kernel Learning defines the following heirarchical model:

$$\begin{aligned}
&\{\text{Hyperprior}\} && p(\boldsymbol{\phi}) = p(\theta, \gamma) \\
&\{\text{Latent GP}\} && g(\omega)|\theta \sim \mathcal{GP}\left(\mu(\omega;\theta), k_g(\omega,\omega';\theta)\right) \\
&\{\text{Spectral Density}\} && S(\omega) = \exp\{g(\omega)\} \\
&\{\text{Data GP}\} && f(x_n)|S(\omega), \gamma \sim \mathcal{GP}(\gamma_0, k(\tau; S(\omega))).
\end{aligned} \tag{5}$$

In the above, $k(\tau; S(\omega))$ is computed using the trapezoid rule of equation 4. We let $f(x)$ be the noiseless outputs of the data, and $y(x)$ be the (potntially) noisy observations: $y(x) \sim \mathcal{GP}(\gamma_0, k(\tau; S(\omega)) + \gamma_1\delta_{\tau=0})$. Full specification of hyperpriors is given in section A.1.

A guiding figure outlining the heirarchy outlined above, showing realizations of the prior of the latent process, kernels constructed from these realizations, and data drawn from GPs with these kernels is given in figure 4 in Section B.

Note that when sampling at N datapoints and W frequencies, the storage costs for this model are naively $\mathbb{O}(N^2 + W^2)$ with the computational costs $\mathbb{O}(N^3 + W^3)$; however, Toeplitz structure in both the data and latent GPs can be exploited to get runtimes of $O(N^3 + I \log I)$ (Guinness & Fuentes, 2017; Wilson et al., 2014).

3. Related Work

Gaussian Process Density Models Leonard (1978); Lenk (1988) extensively studied the logistic transformation of Gaussian processes as a density model, while Tokdar & Ghosh (2007) proved asymptotic properties of the same model as a density estimator. Adams et al. (2009) proposed the Gaussian process density sampler, a generalization of logistic Gaussian processes to include arbitrary non-linearities, proposing a rejection scheme for sampling and inference.

Spectral Domain Gaussian Processes Perhaps the first to consider the Fourier domain of stationary covariance functions was Whittle (1957) who used the spectral density to approximate Gaussian log-likelihoods in linear time. Similarly, Lázaro-Gredilla et al. (2013) proposed using sparse spectral densities as a mechanism for making Gaussian processes more scalable. Wilson & Adams (2013); Wilson (2014) proposed the spectral mixture kernel, fitting a mixture of Gaussians in the spectral domain, exploiting the closed form spectral density for efficient inference.

Alternatively, Rahimi & Recht (2008) proposed approximating Eq. 2 using Monte Carlo samples from the spectral densities, proposing random Fourier features. Oliva et al. (2016) extended the spectral mixture kernel, proposing a non-parametric Bayesian kernel.

Finally, Tobar et al. (2015) proposed a convolutional model for modelling the spectral density with a Gaussian process. For comparison we include predictions generated using the methods of Tobar (2018) in which the spectral density of time series data is modeled nonparametrically using GPs. This method is highly effective for interpolation, but as is shown in Section 5 is not as robust for interpolation tasks as FKL.

4. Inference

For the hierarchical model defined in Equation 5, we need to learn both the hyper-parameters, ϕ, and an instance of the latent Gaussian process, $g(\omega)$. We employ alternating

Algorithm 1 Alternating Sampler

```
Input: Data $(x, y)$, Initial hyper-parameters $\phi_0$, Sampling frequencies $\omega$, Initial Latent GP $g(\omega)$, Number of gradient steps $N_{optim}$, Number of ESS samples $N_{ESS}$,
repeat
    for $i = 1$ to $N_{optim}$ do
        Update $\phi$ using gradient descent given $g(\omega)$ and Eqn. 6
    end for
    for $i = 1$ to $N_{ESS}$ do
        Update $g(\omega)$ using elliptical slice sampling given $\phi$ and Eqn. 7
    end for
until convergence
```

updates in which the latent GP, $g(\omega)$ and the hyperparameters, ϕ are updated separately. A full description of the method is in Algorithm 1.

Updating Hyper-Parameters Considering the model specification in Eq. 5, we can define a loss as a function of ϕ for fixed realization of the latent GP $\tilde{g}(\omega)$ and data observations $y(x)$. The loss corresponds to the prior likelihood, likelihood of the latent realization, and likelihood of the data:

$$\mathcal{L}(\boldsymbol{\theta}) = -\left(\log p(\boldsymbol{\phi}) + \log p(\tilde{g}(\omega)|\theta, \omega) + \log p(y|g(\omega), \gamma, x)\right). \tag{6}$$

We employ the AMSGRAD implementation of Adam as provided by PyTorch (Reddi et al., 2019).

Updating Latent Gaussian Process We perform updates over the latent GP in a fully Bayesian way, noting that the full conditional is

$$p(g(\omega)|\boldsymbol{\phi}, x, y(x), f(x)) \propto \mathcal{N}(\mu(\omega;\theta), k_g(\omega;\theta)) \cdot p(f(x)|g(\omega), \gamma). \tag{7}$$

Elliptical slice sampling (Murray et al., 2010; Murray & Adams, 2010) can be used in this setting to efficiently sample from the latent posterior. In practice, we found that ESS nearly converges in terms of likelihood after just 15-20 iterations.

5. Experiments

In this section we fix the mean and covariance of the latent process to take the following

$$\mu(\omega;\theta) = \theta_0 - \frac{\omega^2}{2\tilde{\theta}_1^{\,2}}$$

$$k_g(\omega, \omega'; \theta) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{|\omega - \omega'|}{\tilde{\theta}_2}\right) K_\nu \left(\sqrt{2\nu}\frac{|\omega - \omega'|}{\tilde{\theta}_2}\right). \tag{8}$$

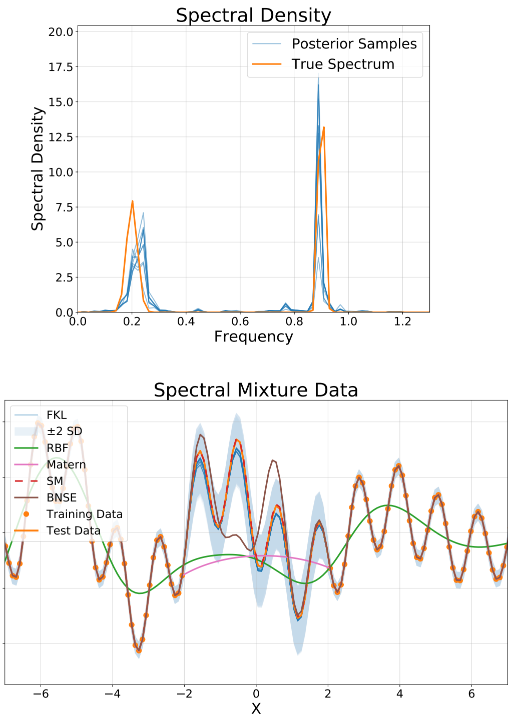

Figure 1. **Above**: Spectrum reconstruction for data generated from a spectral mixture kernel. **Below**: The samples of the latent GP correspond closely to the true spectral density, and many of the FKL predictions on the held out data are nearly on par with the ground-truth model (SM in dashed pink).

The $\tilde{\theta}_i$'s are variables that need to be ensured to be positive, so the above are computed with $\tilde{\theta}_i = \log(e^{\theta_i} + 1)$, the softplus of the raw value.

The mean of the latent process, $\mu(\omega;\theta)$, is (up to some scaling) the log-spectral density of an RBF kernel, giving that our prior mean corresponds to an RBF kernel. The covariance function is a Materǹ with ν chosen to be 1.5. Using a Materǹ kernel to model the log-spectral density allows recovery of sharp peaks that are often seen in the spectrum of covariance functions, in particular when the data contain an overall trend.

Using the sampling routine of Algorithm 1 and transforming the spectral densities into kernels (using the processes outlined in Section 2.2), samples of predictions on the training and testing data can be taken.

In the figures of this section we show spectral densities and predictions using the kernels generated by the final 10 spectral density samples drawn from ESS in the inference procedure. Thus each blue line corresponds to the prediction (with appropriate shading) of one output of the alternating sampler.

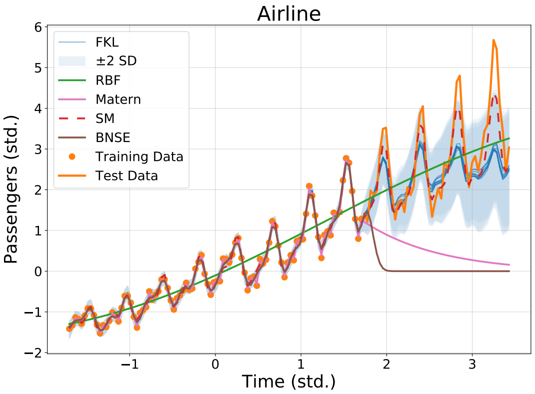

Figure 2. Airline Data

5.1. Recovery of Spectral Mixture Kernels

Synthetic data are generated from a mean zero GP with a spectral mixture (SM) kernel (Wilson & Adams, 2013). The true spectrum corresponds to a mixture of two Gaussians, shown as the orange line in the top panel of figure 1. Using the inference procedure of Section 4 recovery of the true spectrum is obtained. The kernels corresponding to the sampled spectral density and the ground truth are shown in Figure 6 as well as figures from a similar experiment on data drawn from a GP with a quasi-periodic kernel are shown in Section B.

5.2. Extrapolation: Airline Data

We next consider the benchmark airline passenger dataset (Hyndman, 2005) consisting of 96 monthly observations of airline passengers from 1949 to 1961, attempting to extrapolate the next 48 observations. The dataset is considered difficult for zero mean Gaussian processes with stationary kernels due to the absence of noise artifacts, the periodicity, and the linear-like trend over time.

Standard parametric kernels and more modern methods such as Bayesian Nonparametric Spectral Estimation (BNSE) from Tobar (2018) are highly capable of interpolating the training data, but quickly mean-revert in the testing region.

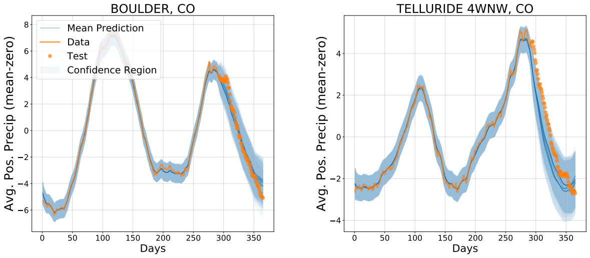

Figure 3. Interpolation and extrpolation on multi-task precipitation data for Boulder, CO and Telluride, CO.

5.3. Multi-task Time Series

FKL provides the framework to consider multi-task data in a novel way. We assume that the covariance function for each task t is constructed from a sample of a shared latent GP over the log-spectral density. Indicating a independent realizations of the latent GP with superscripts ($S^t(\omega) = \exp\{g^t(\omega)\}$) and the GP's over each output dimension of the data with subscripts ($f_t(x)$). Thus the multi-task model is

$$
\begin{aligned}
g(\omega)|\theta &\sim \mathcal{GP}(\mu(\omega;\theta), k_g(\omega,\omega';\theta)) \\
f_t(x)|g^t(\omega),\gamma &\sim \mathcal{GP}(\gamma_0, k(\tau, S^t(\omega)) + \gamma_1\delta_{\tau=0}),
\end{aligned}
\tag{9}
$$

with the same hyperparameters and hyperpriors as Equation 5. The inference procedure is modified only by iterating through tasks and performing elliptical slice sampling updates to each one independently.

We test this average postiive precipitation by day taken from the United States Historical Climatology network (Menne et al., 2015). We look at modeling data from climatologically similar recording stations in Colorado, using the last 2 months of the year as held-out testing data. Results are shown for two such stations in figure 3.

6. Discussion

Functional kernel learning (FKL) permits not only accurate reconstructions of spectral densities corresponding to stationary kernels, but also extrapolates and interpolates observable data via Fourier transformation and latent Gaussian processes. In turn, this permits the inference of spectral densities and kernels in an analytic and probabilistic manner. Learning from data incorporates gradient descent and elliptical slice sampling in order to perform updates of the latent GP and sampled spectral densities. The method is validated in recovery of known kernels (and corresponding spectral densities) as well as accurate interpolation and extrapolation of both synthetic and real-world data.

Future directions of work include: (i) discovering shared structure across multiple heterogenous tasks; (ii) exploiting structure inherent in stationary kernels, such as Toeplitz structure, for increased scalability; (iii) exploring generalizations of FKL corresponding to learning non-stationary kernels.

References

Adams, R. P., Murray, I., and MacKay, D. J. C. Nonparametric Bayesian Density Modeling with Gaussian Processes. *arXiv:0912.4896 [math, stat]*, December 2009. URL http://arxiv.org/abs/0912.4896. arXiv: 0912.4896.

Bochner, S. *Lectures on Fourier integrals*. Princeton University Press, 1959.

Damianou, A. and Lawrence, N. Deep gaussian processes. In *Artificial Intelligence and Statistics*, pp. 207–215, 2013.

Genton, M. G. Classes of kernels for machine learning: a statistics perspective. *Journal of machine learning research*, 2(Dec):299–312, 2001.

Guinness, J. and Fuentes, M. Circulant Embedding of Approximate Covariances for Inference From Gaussian Data on Large Lattices. *Journal of Computational and Graphical Statistics*, 26(1):88–97, January 2017. ISSN 1061-8600, 1537-2715. doi: 10.1080/10618600.2016.1164534. URL https://www.tandfonline.com/doi/full/10.1080/10618600.2016.1164534.

Hyndman, R. J. Time series data library, 2005. URL http://www-personal.buseco.monash.edu.au/~hyndman/TSDL/.

Lenk, P. J. The logistic normal distribution for bayesian, nonparametric, predictive densities. *Journal of the American Statistical Association*, 83(402):509–516, 1988.

Leonard, T. Density estimation, stochastic processes and prior information. *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(2):113–132, 1978.

Lázaro-Gredilla, M., Quiñonero-Candela, J., Rasmussen, C. E., and Figueiras-Vidal, A. R. Sparse Spectrum Gaussian Process Regression. pp. 17, 2013.

Menne, M. J., Williams, C. N., and Vose, R. S. United states historical climatology network daily temperature, precipitation, and snow data. *Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, Oak Ridge, Tennessee*, 2015.

Murray, I. and Adams, R. P. Slice sampling covariance hyperparameters of latent gaussian models. In *Advances in neural information processing systems*, pp. 1732–1740, 2010.

Murray, I., Prescott Adams, R., and MacKay, D. J. Elliptical slice sampling. In *Artificial Intelligence and Statistics*, 2010.

Oliva, J. B., Dubey, A., Wilson, A. G., Póczos, B., Schneider, J., and Xing, E. P. Bayesian nonparametric kernel-learning. In *Artificial Intelligence and Statistics*, pp. 1078–1086, 2016.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pp. 1177–1184, 2008.

Reddi, S. J., Kale, S., and Kumar, S. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.

Remes, S., Heinonen, M., and Kaski, S. Non-Stationary Spectral Kernels. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 4642–4651. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7050-non-stationary-spectral-kernels.pdf.

Tobar, F. Bayesian nonparametric spectral estimation. In *Advances in Neural Information Processing Systems*, pp. 10127–10137, 2018.

Tobar, F., Bui, T. D., and Turner, R. E. Learning stationary time series using gaussian processes with nonparametric kernels. In *Advances in Neural Information Processing Systems*, pp. 3501–3509, 2015.

Tokdar, S. T. and Ghosh, J. K. Posterior consistency of logistic gaussian process priors in density estimation. *Journal of statistical planning and inference*, 137(1):34–42, 2007.

Whittle, P. On the use of the normal approximation in the treatment of stochastic processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 19(2):268–281, 1957.

Williams, C. K. and Rasmussen, C. E. *Gaussian processes for machine learning*, volume 2. MIT Press Cambridge, MA, 2006.

Wilson, A. and Adams, R. Gaussian process kernels for pattern discovery and extrapolation. In *International Conference on Machine Learning*, pp. 1067–1075, 2013.

Wilson, A. G. *Covariance kernels for fast automatic pattern discovery and extrapolation with Gaussian processes*. PhD thesis, 2014.

Wilson, A. G., Gilboa, E., Nehorai, A., and Cunningham, J. P. Fast kernel learning for multidimensional pattern extrapolation. In *Advances in Neural Information Processing Systems*, pp. 3626–3634, 2014.

Wilson, A. G., Hu, Z., Salakhutdinov, R., and Xing, E. P. Deep kernel learning. In *Artificial Intelligence and Statistics*, pp. 370–378, 2016.

A. Extensions

Multi-Output Hierarchical Gaussian Processes Given multiple outputs, $y_1(x_1)$, that might be correlated, it might be natural to consider that the spectral densities themselves are related by the same Gaussian process (e.g. have the same parameters $\boldsymbol{\theta}$) with only the specific random function draw (the specific $g(s|\boldsymbol{\theta})$) being different).

Deep Gaussian Process The model specification is implicitly a *deep* Gaussian process, albeit one with the depth occurring only in the covariance structure. Here, the recursion is in contrast to the Damianou & Lawrence (2013)'s construction of deep Gaussian processes as compositions of standard Gaussian processes on potentially different inputs; by contrast, our formulation allows for exact inference without the learning of inducing points or pseudo-inputs.

Non-Stationary Kernels Despite being the most generally employed family of kernels, monotonic stationary kernels such as the Gaussian kernel are not only a subpar choice for partitioning input spaces, but also a special case of the non-stationary class of kernels (Genton, 2001). Non-stationary kernels are often employed in signal processing, geostatistics, and time-series analysis. Unlike the Gaussian kernel, one such member of the partly non-stationary class includes a variant of the Spectral Mixture kernel (Wilson, 2014), a covariance function often used in modeling physical processes due to its construction of a univariate spectral density by a mixture of normal distributions. Combining standard kernels with various transformations (Wilson et al., 2016) with products of stationary kernels has been a common kernel construction approach. Perhaps the simplest non-stationary kernel is the dot product kernel as a way to model input-dependent variance (Williams & Rasmussen, 2006). Despite being unsuited for modeling non-monotonic properties, the aforementioned non-stationary kernels are useful for modeling dynamical systems.

A.1. Prior Specification

For the noise terms, we place smoothed box priors on the range (1e-8, 1e-3) to control both numerical instability and the noise terms. A smooth box prior is a smooth approximation to uniform priors, where $B(x) = \{a \leq x \leq b\}$ then $d(x, B) := \min_{x' \in B} |x - x'|$ and finally the density is given by $f(x) := \exp\{-d(x, B)^2/\sqrt{2\sigma^2}\}$. See `https://gpytorch.readthedocs.io/en/latest/priors.html` for further implementation details. For the constant mean terms in both the data and latent means, we place uninformative $\mathcal{N}(0, 100)$ priors to allow broad ranges of outcomes. For the length-scale in the spectral density mean along with the length-scale and output-scale of the covariance of the spectral density GP, we place standard log-normal priors, as these variables need to be positive.

B. Additional Figures

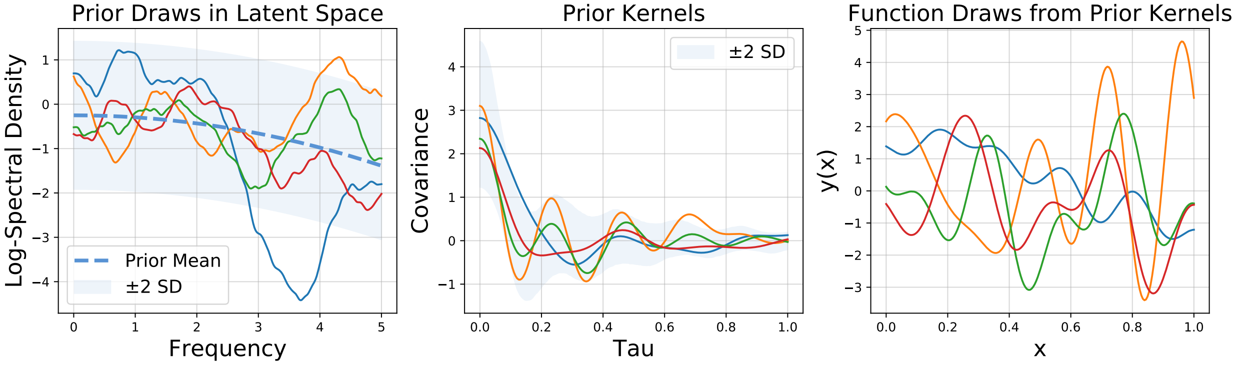

Figure 4. **Left**: Using randomly initialized hyper-parameters functions are drawn from the latent GP. **Center**: The latent realizations are used to compose kernels over data. **Right**: Using each of the constructed kernels mean-zero functions over data are drawn. Shaded regions show 2 standard deviations above and below the mean.

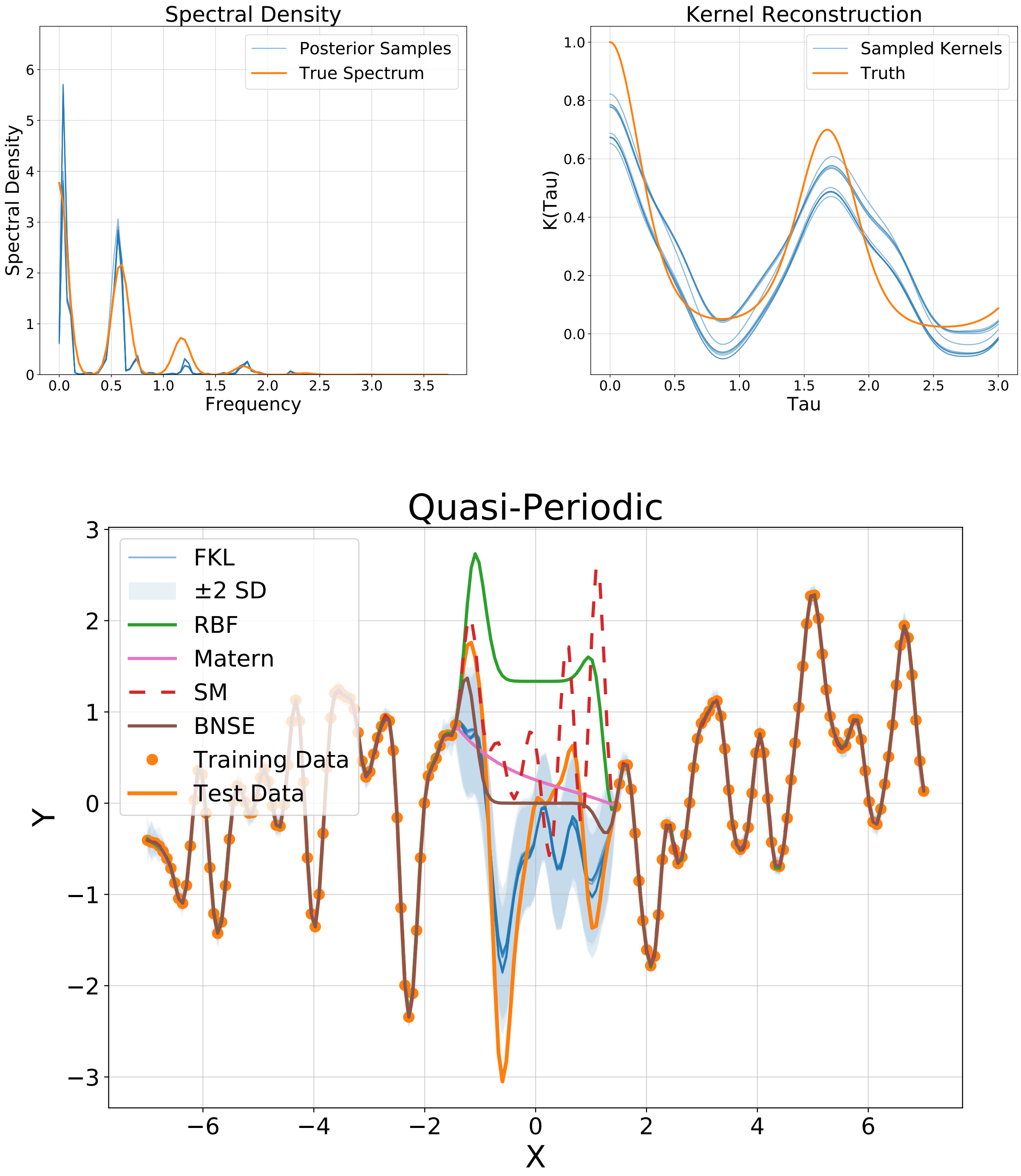

Figure 5. **Above Left:** Spectrum reconstruction for a Quasi-periodic kernel; **Above Right:** Kernel reconstruction; **Below:** Data interpolation and extrapolation for FKL and competing methods

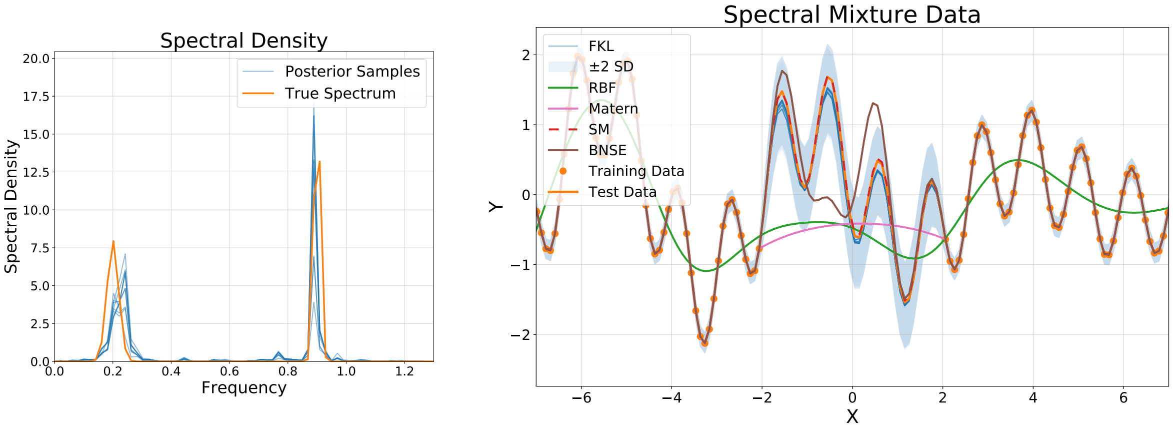

Figure 6. **Left**: Spectrum and kernel reconstruction for data generated from a spectral mixture kernel. **Right**: The samples of the latent GP correspond closely to the true spectral density, and many of the FKL predictions on the held out data are nearly on par with the ground-truth model (SM in dashed pink).