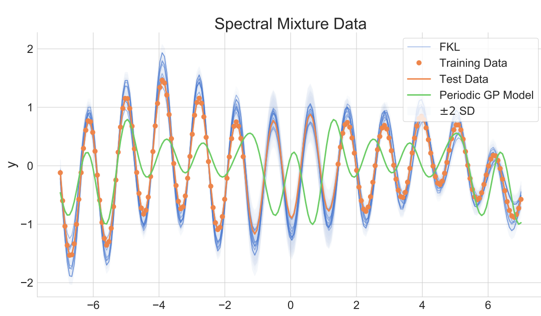
We thank all the reviewers for their supportive and insightful comments. While kernel learning has now been broadly identified as important for good performance, the vast majority of approaches, while highly useful, focus on parametric methods that do not represent uncertainty over the values of the kernel, can be difficult to train, and difficult to specify inductive biases. Our proposed functional kernel learning (FKL) approach provides a Bayesian nonparametric distribution over kernels, with (a) support for a wide range of kernels; (b) uncertainty representation; (c) easy specification of inductive biases through prior means on the distribution over the spectral density; (d) automatic inference without requiring extensive intervention; (e) natural multidimensional and multi-output generalizations; (f) exhaustive experimental results over a wide range of problems supporting the procedure. Moreover, we want to emphasize that the approach is broadly applicable, and does well on data with and without periodic structure. We would be grateful if reviewers could consider our response in determining their ultimate assessment.

R2: *Bayesian Linear Regression:* Thank you for pointing out the theoretically and practically useful connection to Bayesian linear regression with trigonometric basis functions. When we use a transformed Gaussian process prior on the spectral density with a Matern-3/2 kernel, which is mean square continuous, we make the assumption that $S(\omega)$ is a continuous density function – which we will clarify in the camera ready version. However, one can make other choices of covariance function for the GP on the spectrum. To show density amongst spectral measures, we can adapt Theorem 5 of [3] to our setting, noting that the trapezoid rule can be shown to be equivalent to both Riemann and Darboux sums. For discontinuous but finite measures the trapezoid rule will provide an approximator of an underlying stationary kernel on the compact set $[0, \omega_{max}]$, converging as $\omega_{max} \to \infty$ (e.g. as the number of basis functions goes to infinity), where the trapezoids can represent mixtures of Gaussians on the spectral density, and Gaussian mixtures are dense approximations of Riemann integrable densities [3] (by collapsing onto point masses). For continuous spectral densities, we note that in practice the rate of convergence would typically be much faster than a standard Fourier series approximation corresponding to point masses on the spectrum.

Latent Mean function: We choose a quadratic mean function for the GP on the spectral density to induce a prior expectation of an RBF kernel (line 115). A major and distinctive advantage of the FKL model is the ability to specify the prior distribution over kernel classes, which can provide a powerful inductive bias. FKL with a quadratic mean will have the inductive biases of an RBF kernel, but has the ability to respond to patterns in the data to learn non-RBF covariance structures. *Higher dimensions*: We will revise the writing of this section to emphasize a) the limitations of product kernels in high dimensions and b) the need to use multivariate FFTs to fully represent multi-dimensional stationary functions. *Comparisons with SM kernels on UCI*: Inspired by your feedback, we ran several experiments with exact product SM kernels, which performed somewhat worse than FKL (and require a lot more manual intervention and careful initialization). The RMSEs for product SM kernels are shown below.

fertility	concreteslump	servo	machine	yacht	housing	energy
0.199 ± 0.038	63.857 ± 8.111	0.276 ± 0.117	1.024 ± 0.181	0.224 ± 0.083	9.54 ± 1.668	9.681 ± 16.62

R1: Thanks for your helpful comments. In the camera ready, we will fix the typos and add in-text references to the figures we missed. *1,2*: For non-stationary kernels, we will discuss how we could extend to the multivariate generalized Fourier transform [1,2]. Non-axis aligned methods are also possible with other generalizations of FFT (possibly [3]). *3*: Yes, this means that scale factors are unnecessary. *5*: We will clarify that Δ is chosen to be no greater than the maximum space between any adjacent points along any dimension of the data. *7*: We standardize and de-mean the number of passengers per month in 1000s. In the camera ready, we will update the figure to be on the count instead. *9*: FKL in its current form is most natural for continuous outputs, but could be adapted (with an appropriate likelihood) for ordinal or categorical data.



R4: Thanks for your thoughtful comments. *Time scaling properties in increased dimensions:* We would like to clarify that the product kernel for multiple dimensions separates as a product of one dimensional integration problems allowing the computational complexity to scale linearly with dimension. *Comparisons with periodic warped GPs:* We have tried comparisons to GPs with periodic kernels. Inspired by your feedback, we ran the experiment with associated figure on the left, where the data are drawn from a GP with spectral mixture kernel. We see FKL outperforms a standard GP with a periodic kernel. We also note that many of the UCI datasets do not have quasi-periodic behaviour and we still see FKL outperforming standard kernels; the benefit of FKL is that the kernel is not restricted to a particular functional form and can learn many types of stationary kernels.

References:

[1] Samo and Roberts, Generalized Spectral Mixture Kernels, https://arxiv.org/abs/1506.02236
[2] Remes et al, Generalized Spectral Mixture Kernels, NeurIPS, 2017.
[3] Shen, Z., Heinonen, M., and Kaski, S. Harmonizable mixture kernels with variational Fourier features. AISTATS, 2019.