# Function-Space Distributions over Kernels

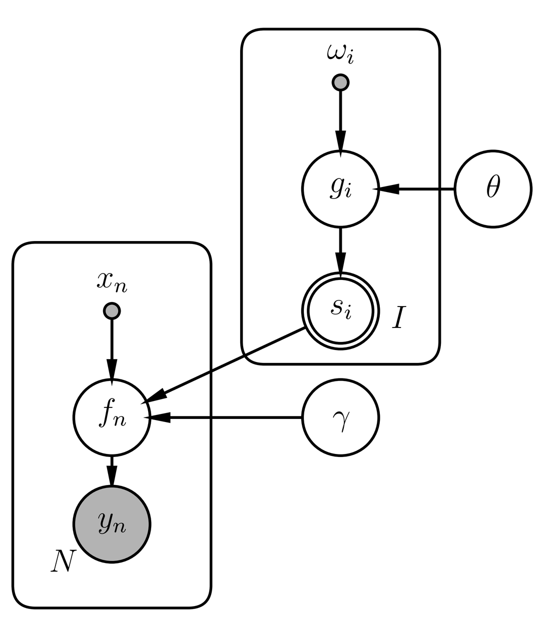
Greg Benton[1], Wesley Maddox[1], Jayson Salkey[1], Julio Albinati[2], Andrew Gordon Wilson[1]

[1] Cornell University, [2] Microsoft

## Outline

- Gaussian processes (GPs) are powerful regression tools, relying only on mean and covariance functions
- In general can assume constant mean; care most about uncovering the covariance through *kernel learning*
- Learn everything about a stationary covariance function $k(\tau)$ by learning its spectral density $S(\omega)$
- Propose *Functional Kernel Learning* (FKL): model the log-spectral density of a kernel using a latent GP
- *Representation Learning* with Gaussian processes through the construction of unconventional kernels

## Spectral Representation



**Bochner's Theorem:** Can represent a (positive definite) kernel as

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

with $\tau = |x - x'|$, the distance between any two inputs, and some positive finite function $S(\omega)$. Only need to learn the (unnormalized) spectral density $S(\omega)$ to learn $k(\tau)$.

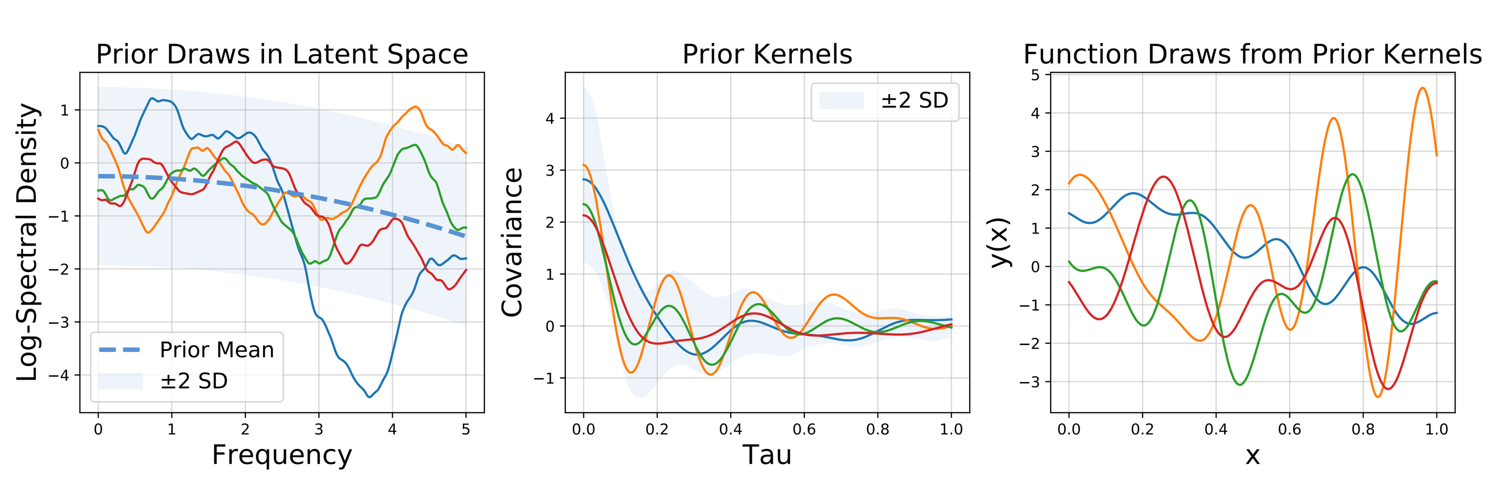
**Model Specification**

$$\{\text{Hyperprior}\} \quad p(\boldsymbol{\phi}) = p(\theta, \gamma)$$

$$\{\text{Latent GP}\} \quad g(\omega)|\theta \sim \mathcal{GP}\left(\mu(\omega;\theta), k_g(\omega, \omega'; \theta)\right)$$

$$\{\text{Spectral Density}\} \quad S(\omega) = \exp\{g(\omega)\}$$

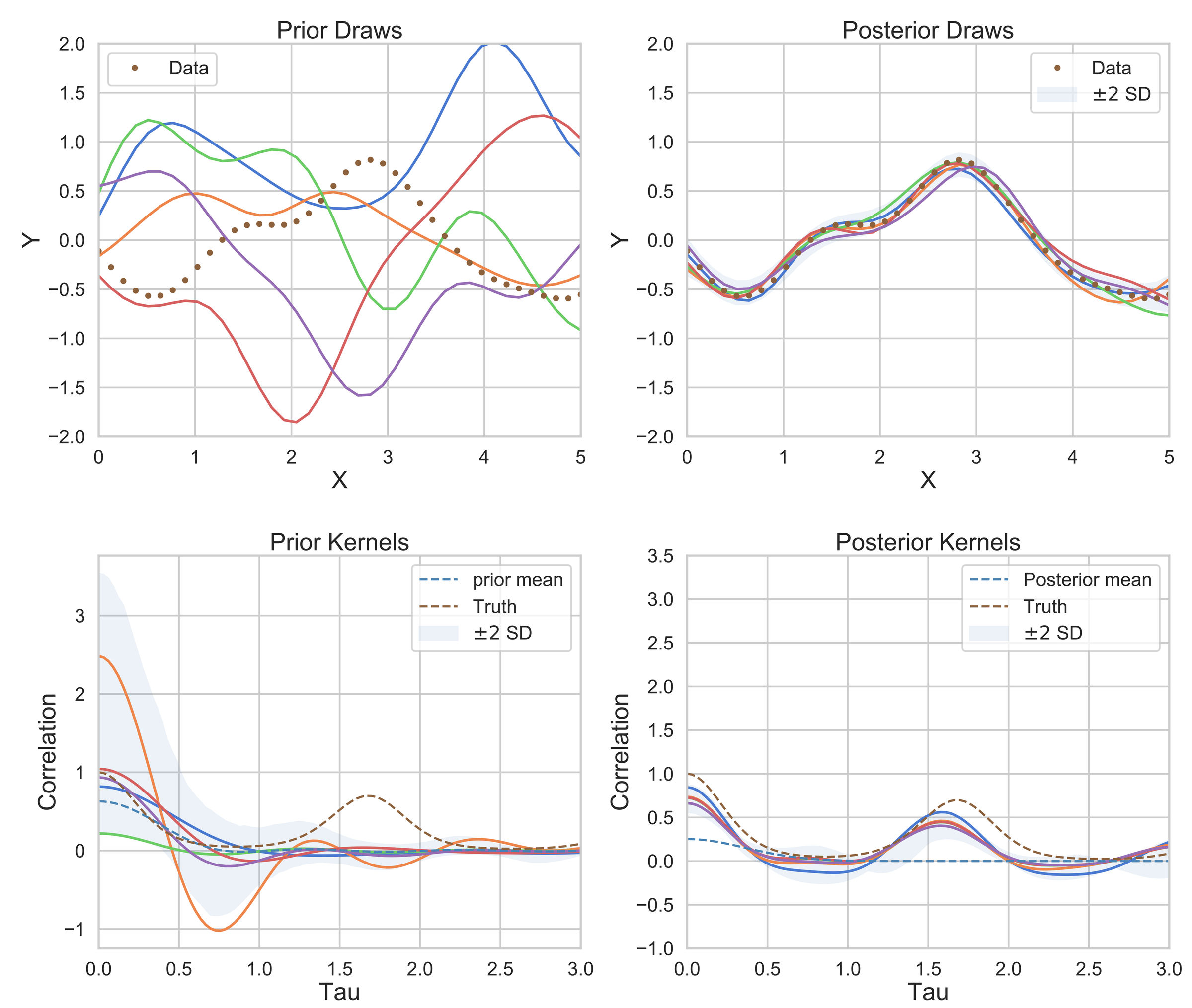
$$\{\text{Data GP}\} \quad f(x_n)|S(\omega), \gamma \sim \mathcal{GP}(\gamma_0, k(\tau; S(\omega))).$$




**Inference:** We employ Monte Carlo EM to learn the hyper-parameters and instance of the spectral density corresponding to the kernel over data $k(\tau)$:

- Initialize $g(\omega)$ to log-periodogram and $\phi$ to likely values from $p(\boldsymbol{\phi})$ then repeat:
  - Fix $g(\omega)$ and update $\boldsymbol{\phi}$ using gradient descent
  - Fix $\boldsymbol{\phi}$ and update $g(\omega)$ using elliptical slice sampling

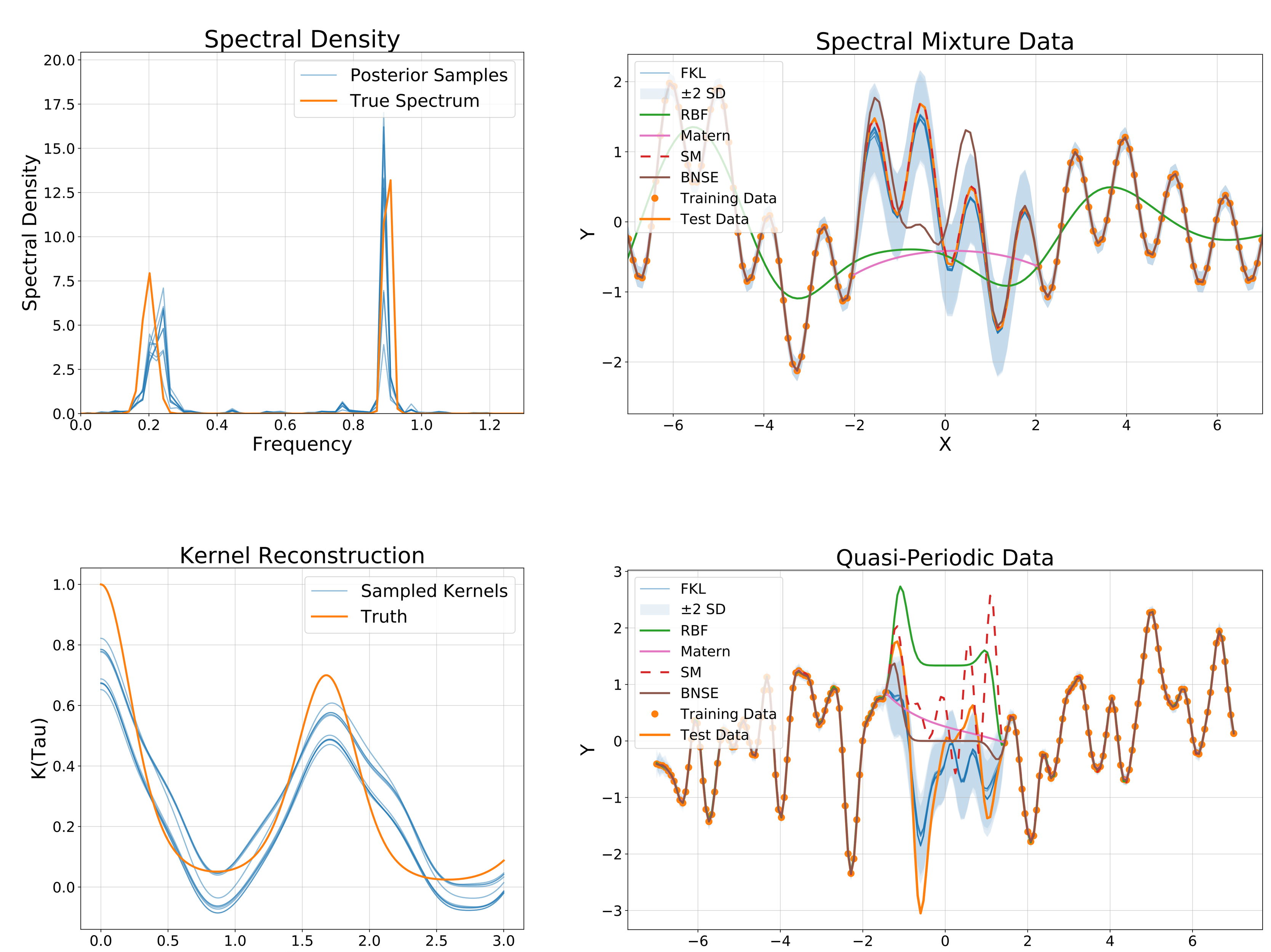
## Functional Kernel Learning



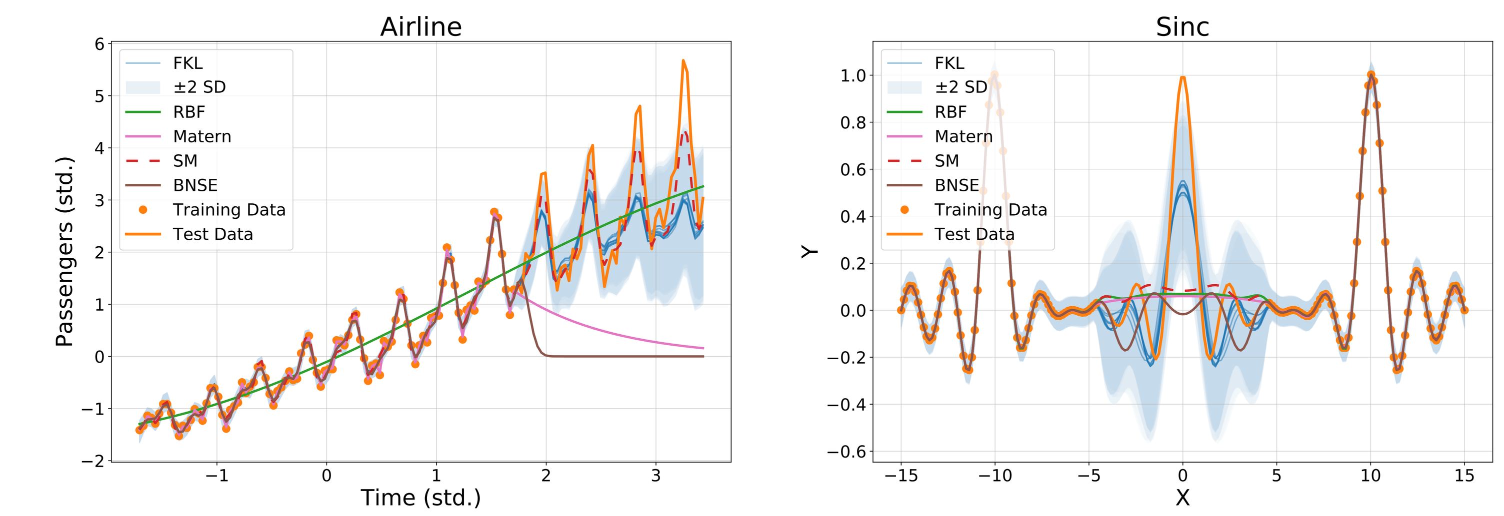

- Top plots show standard GP regression process over data
- Bottom plots show similar process over *kernels*
- FKL provides non-parametric regression interpretation of *kernel learning*

## Recovering Spectra of Known Kernels

- Generate data from GP with known kernel: spectral density is mixture of two Gaussians
- Learn true (known) spectral density and kernel using alternating inference strategy




## Extrapolation and Interpolation




## Multi-Task Learning

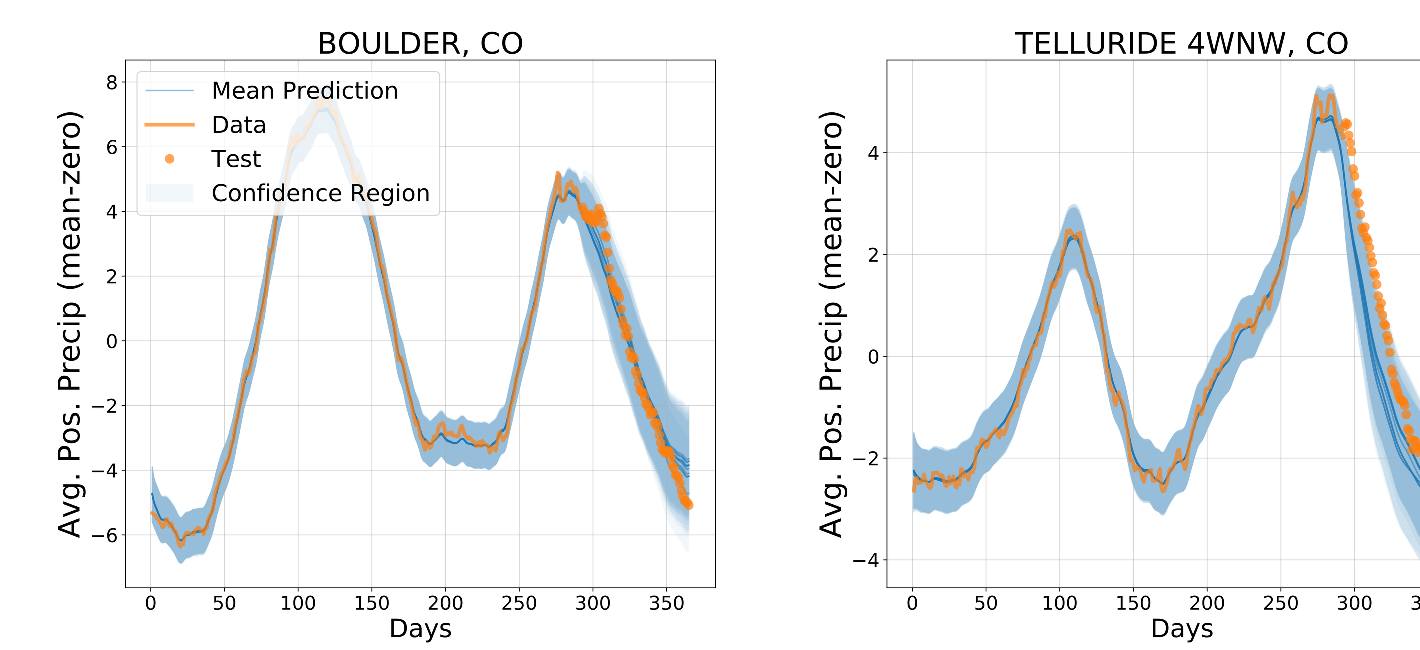
Tie covariance for each location together by assuming single latent GP over log-spectral density. With $g^t(\omega)$ as the $t^{th}$ realization of $g(\omega)$ we have,

$$\{\text{Latent GP}\} \quad g(\omega)|\theta \sim \mathcal{GP}\left(\mu(\omega;\theta), k_g(\omega, \omega'; \theta)\right)$$

$$\{\text{Task GP}, \forall t\} \quad f_t(x)|g^t(\omega), \gamma \sim \mathcal{GP}(\gamma_0, k(\tau; S(\omega)))$$

where the hyperprior and link functions are the same as in the single dimensional case.

- Multi-output time series data from US Historical Climatology Network:
  - Average positive precipitation by day
  - Can run extrapolation simulations on >100 stations (N = 40k) at a time




## Results

- Developed FKL: a non-parametric approach to kernel learning
- Outperforms state-of-the-art on extrapolation while requiring less intervention
- Extends to multi-task data in a new and intuitive way via linking spectral densities