FUNCTION SPACE DISTRIBUTIONS OVER KERNELS

GREG BENTON, WESLEY MADDOX, JAYSON SALKEY, JULIO ALBINATI, ANDREW GORDON WILSON
HIGH LEVEL IDEA

- Gaussian Process (GP): stochastic process for which any finite collection of points is jointly normal

- $k(x, x')$ a kernel function describing covariance

$$y(x) \sim GP(\mu(x), k(x, x'))$$
**HIGH LEVEL IDEA**

\[ y(x) \sim GP(\mu(x), k(x, x')) \]
OUTLINE

- Introduction
  - Mathematical Foundation
  - Model Specification
  - Inference Procedure
FUNCTIONAL KERNEL LEARNING

OUTLINE

‣ Introduction

‣ Experimental Results
  ‣ Recovery of known kernels
  ‣ Interpolation and extrapolation of real data
OUTLINE

- Introduction
- Experimental Results
- Extension to multi-task time-series
  - Precipitation data
BOCHNER’S THEOREM

- If $k(x, x') = k(\tau)$ then we can represent $k(\tau)$ via its spectral density:

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

- Learning the spectral representation of $k(\tau)$ is sufficient to learn the entire kernel
BOCHNER’S THEOREM

- If \( k(x, x') = k(\tau) \) then we can represent \( k(\tau) \) via its spectral density:
  \[
  k(\tau) = \int_{\mathbb{R}} e^{2\pi i \omega \tau} S(\omega) d\omega
  \]
- Learning the spectral representation of \( k(\tau) \) is sufficient to learn the entire kernel
- Assuming \( k(\tau) \) is symmetric and data are finitely sampled, the reconstruction simplifies to:
  \[
  k(\tau) = \int_{[0, \pi/\Delta)} \cos(2\pi \tau \omega) S(\omega) d\omega
  \]
FUNCTIONAL KERNEL LEARNING

Graphical Model

\[
p(\phi) = p(\theta, \gamma)
\]

Hyper-prior

\[
g(\omega) \mid \theta \sim GP\left(\mu(\omega; \theta), k_g(\omega, \omega'; \theta)\right)
\]

Latent GP

\[
S(\omega) = \exp\{g(\omega)\}
\]

Spectral Density

\[
f(x) \mid S(\omega), \gamma \sim GP(\gamma_0, k(\tau; S(\omega)))
\]

Data GP
\[ p(\phi) = p(\theta, \gamma) \]

\[ g(\omega) | \theta \sim \text{GP} \left( \mu(\omega; \theta), k_g(\omega, \omega'; \theta) \right) \]

\[ S(\omega) = \exp\{g(\omega)\} \]

\[ f(x) | S(\omega), \gamma \sim \text{GP}(\gamma_0, k(\tau; S(\omega)) + \gamma_1 \delta_{\tau=0}) \]
LATENT MODEL

Mean of latent GP is log of RBF spectral density

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

Covariance is Matérn with $\nu = 1.5$

$$k_g(\omega, \omega'; \theta) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left( \sqrt{2\nu} \frac{|\omega - \omega'|}{\tilde{\theta}_2} \right) K_\nu \left( \sqrt{2\nu} \frac{|\omega - \omega'|}{\tilde{\theta}_2} \right) + \tilde{\theta}_3 \delta_{\tau=0}$$

$$\tilde{\theta}_i = \text{softmax}(\theta_i)$$
FUNCTIONAL KERNEL LEARNING

INERENCE

- Need to update the hyper parameters \( \phi \) and the latent GP \( g(\omega) \)
- Initialize \( g(\omega) \) to the log-periodogram of the data
- Alternate:
  - Fix \( g(\omega) \) and use Adam to update \( \phi \)
  - Fix \( \phi \) and use elliptical slice sampling to draw samples of \( g(\omega) \)
OUTLINE

- Introduction
- Experimental Results
  - Recovery of known kernels
  - Interpolation and extrapolation of real data
DATA FROM A SPECTRAL MIXTURE KERNEL

- Generative kernel has mixture of Gaussians as spectral density
DATA FROM A SPECTRAL MIXTURE KERNEL

Spectral Mixture Data

- FKL
- ±2 SD
- RBF
- Matern
- SM
- BNSE
- Training Data
- Test Data
AIRLINE PASSENGER DATA

FUNCTIONAL KERNEL LEARNING

The graph shows the passenger data over time, with different kernel functions (FKL, RBF, Matern, SM, BNSE) represented by various lines. The data is segmented into training and test data, indicated by different markers. The y-axis represents the standardized number of passengers, and the x-axis represents time in standard deviations.
OUTLINE

- Introduction
- Experimental Results
- Extension to multi-task time-series
  - Precipitation data
Can ‘link’ multiple time series by sharing the latent GP across outputs.

Let \( g^t(\omega) \) denote the \( t^{th} \) realization of the latent GP and \( f_t(x) \) be the GP over the \( t^{th} \) time-series.

Hyper-prior: \( p(\phi) = p(\theta, \gamma) \)

Latent GP: \( g(\omega) | \theta \sim GP\left(\mu(\omega; \theta), k_g(\omega, \omega'; \theta)\right) \)

\( t^{th} \) Spectral Density: \( S^t(\omega) = \exp\{g^t(\omega)\} \)

GP for \( t^{th} \) task: \( f_t(x) | S(\omega), \gamma \sim GP(\gamma_0, k(\tau, S^t(\omega))) + \gamma_1 \delta_{\tau=0} \)
MULTIPLE TIME SERIES

- Can ‘link’ multiple time series by sharing the latent GP across outputs

- Let $g^t(\omega)$ denote the $t^{th}$ realization of the latent GP and $f_t(x)$ be the GP over the $t^{th}$ time-series

  - Hyper-prior: $p(\phi) = p(\theta, \gamma)$
  - Latent GP: $g(\omega) \mid \theta \sim GP\left(\mu(\omega; \theta), k_g(\omega, \omega'; \theta)\right)$
  - $t^{th}$ Spectral Density: $S^t(\omega) = \exp\{g^t(\omega)\}$
  - GP for $t^{th}$ task: $f_t(x) \mid S(\omega), \gamma \sim GP(\gamma_0, k(\tau; S^t(\omega)) + \gamma_1 \delta_{\tau=0})$

- Test this on data from USHCN, daily precipitation values from continental US

  - Inductive bias: yearly precipitation for climatologically similar regions should have similar covariance, similar spectral densities
FUNCTIONAL KERNEL LEARNING

PRECIPITATION DATA

Ran on two climatologically similar locations

BOULDER, CO

TELLURIDE 4WNW, CO

Avg. Pos. Precip (mean-zero)

Days

Avg. Pos. Precip (mean-zero)

Days
PREcipitation Data

Used 108 locations across the Northeast USA

Each station, n = 300

Total: 300 * 108 = 32,400 data points
FKL: Nonparametric, function-space view of kernel learning

Can express any stationary kernel with uncertainty representation

GPY Torch Code: [https://github.com/wjmaddox/spectralgp](https://github.com/wjmaddox/spectralgp)
FUNCTIONAL KERNEL LEARNING

CONCLUSION

‣ FKL: Nonparametric, function-space view of kernel learning

‣ Can express any stationary kernel with uncertainty representation

‣ GPyTorch Code: https://github.com/wjmaddox/spectralgp

QUESTIONS? 

‣ Poster 52
REFERENCES


FUNCTIONAL KERNEL LEARNING

SINC DATA

\[ \text{sinc}(x) = \frac{\sin(\pi x)}{\pi x} \]
Generative kernel is product of RBF and periodic kernels.
QUASI-PERIODIC DATA

- Generative kernel is product of RBF and periodic kernels
ELLiptical SLice Sampling (Murray, Adams, Mackay, 2010)

Sample zero mean Gaussians

Re-parameterize for non-zero mean

| Input: | current state $f$, a routine that samples from $\mathcal{N}(0, \Sigma)$, log-likelihood function $\log L$. |
| Output: | a new state $f'$. When $f$ is drawn from $p^*(f) \propto \mathcal{N}(f; 0, \Sigma) L(f)$, the marginal distribution of $f'$ is also $p^*$. |

1. Choose ellipse: $\nu \sim \mathcal{N}(0, \Sigma)$
2. Log-likelihood threshold:
   $$ u \sim \text{Uniform}[0, 1] $$
   $$ \log y \leftarrow \log L(f) + \log u $$
3. Draw an initial proposal, also defining a bracket:
   $$ \theta \sim \text{Uniform}[0, 2\pi] $$
   $$ [\theta_{\min}, \theta_{\max}] \leftarrow [\theta - 2\pi, \theta] $$
4. $f' \leftarrow f \cos \theta + \nu \sin \theta$
5. if $\log L(f') > \log y$ then:
6. Accept: return $f'$
7. else:
   Shrink the bracket and try a new point:
8. if $\theta < 0$ then: $\theta_{\min} \leftarrow \theta$ else: $\theta_{\max} \leftarrow \theta$
9. $\theta \sim \text{Uniform}[\theta_{\min}, \theta_{\max}]$
10. GoTo 4.

Figure 2: The elliptical slice sampling algorithm.

Figure 3: (a) The algorithm receives $f=\times$ as input. Step 1 draws auxiliary variate $\nu=+$, defining an ellipse centred at the origin (a). Step 2: a likelihood threshold defines the ‘slice’ (—). Step 3: an initial proposal (●) is drawn, in this case not on the slice. (b) The first proposal defined both edges of the $[\theta_{\min}, \theta_{\max}]$ bracket; the second proposal (●) is also drawn from the whole range. (c) One edge of the bracket (——) is moved to the last rejected point such that $\times$ is still included. Proposals are made with this shrinking rule until one lands on the slice. (d) The proposal here (●) is on the slice and is returned as $f'$. (e) Shows the reverse configuration discussed in Section 2.3: $\times$ is the input $f'$, which with auxiliary $\nu'=+$ defines the same ellipse. The brackets and first three proposals (●) are the same. The final proposal (●) is accepted, a move back to $f$. 