Scalable Gaussian Processes for Characterizing Multidimensional Change Surfaces

William Herlands¹, Andrew Gordon Wilson ², Hannes Nickisch ², Seth Flaxman ³, Daniel Neill ¹, Wilbert van Panhuis ¹, Eric Xing ¹

¹Carnegie Mellon University, ²Philips Research Hamburg, ³University of Oxford, ⁴University of Pittsburgh

Introduction

We introduce a GP model that is capable of automatically learning expressive covariance functions, including a sophisticated continuous change surface. We derive scalable inference procedures leveraging Kronecker structure and a lower bound on the marginal likelihood using the Weyl inequality. As compared to previous change point models, our approach allows accurate modeling and prediction for complex changes often observed in human data that are multidimensional, gradual, and heterogeneous.

Gaussian Processes

A Gaussian process is a nonparametric prior over functions

\[ f(x) \sim GP(m(x), k(x, x')) \]

We define the covariance function \( k(x, x') = k(x, x') + \epsilon \)

Any function \( f(x) \) that satisfies \( f(x) \mid N(0, k(x, x')) \) can be expressed as a Gaussian process, \( k(x, x') = k(x, x') + \epsilon \).

Marginal Likelihood

\[ \log \pi(x) = -\log \det \Sigma - \frac{1}{2} x^\top \Sigma^{-1} x + \epsilon \]

Weighting functions \( w(x) \)

The expression \( w(x) \) is determined by how changes occur in the data, and how many changes occur in our data. We do not require any prior knowledge about the functional form of \( w(x) \) and instead assume a Gaussian process prior on \( w(x) \). We approximate the Gaussian process with random Kink Stump features.

Design choices for \( K \)

Each latent function is specified by a kernel with unique hyperparameters. In order to maintain maximum generality, the model uses spectral mixture kernels where \( k_{ij}(x, x') = \sum_{m=1}^{M} \omega_m \cos(2\pi \langle x \rangle - \langle x' \rangle) \], where \( i \in [R^D] \) and \( \Sigma_m = diag(\omega_m) \) is a diagonal covariance matrix for multidimensional inputs.

Nonstationary additive kernel

If we assume independent GP priors on each latent function, we can define \( y(x) = f(x) + z \) where \( f(x) \) has a Gaussian process prior with covariance function,

\[ k(x, x') = \sum_{i=1}^{D} k_i(x, x') + \epsilon \]

\( \sigma_i(x) \) induce nonstationarity since they are dependent on the input \( x \). Thus, even if we use stationary kernels for all \( k_i \), our model results additive and nonstationary.

Acknowledgments

This material is based upon work supported by the NSF Graduate Research Fellowship under Grant No. DGE 1252522 and the NSF award No. IIS-0953300. Flaxman was supported by EPSRC (EP/K009362/1).

Scalable inference

Kronecker methods for grid data

Analytic inference requires the log marginal likelihood (Eq. 4). This involves costly computation of \( \sum_{i=1}^{D} \log(\sigma_i(x)^2) \). Kronecker methods decompose the covariance matrix \( K = \sum_{i=1}^{D} K_i \), where each \( K_i \) is \( n \times n \) such that \( \sum_{i=1}^{D} K_i = K \).

Log determinant \( \log(|K|) \): Weyl's inequality states that for \( n \times n \) Hermitian matrices, \( M = A + B \), with sorted eigenvalues \( \lambda_1, \ldots, \lambda_n \), respectively,

\[ \lambda_i + \lambda_j \leq \lambda_i^J \]

The log determinant \( \log(|K|) \)

\[ \log(|K|) + \log(|A + B|) = \sum_{i,j=1}^{n} \log(|A_i + B_j|) \]

We iteratively apply this approximation to pairs of matrices in order to bound \( \log(\det K) \) to a series of smaller matrices.

Additive Kronecker approximation

Inverse \( -1 \): use finite difference methods to compute linear conjugate gradients. The key subroutine in MVM so the sum of Kronecker products can be efficiently multiplied by a vector.

\[ \sum_{i=1}^{D} \log(\sigma_i(x)) \]

Figure 1: Left shows the time to compute each approximation and the truth. Right shows the time to compute each approximation and the truth.

Scaled additive kernels

Rewrite Eq. 9 in matrix notation where \( \Sigma_i = \text{diag}(\omega_i(x)) \)

\[ S_i = \text{diag}(\sigma_i(x)^2)) \]

Employing the bound on eigenvalues of matrix products,

\[ \log(|A + B|) \leq \log(|A|) + \log(|B|) \]

we can bound \( \log(|K|) \) in Eq. 13 with a Weyl approximation over \( (k_{ij} + \sigma_i(x)^2 + \omega_i(x)) \), where \( n_i \) is the \( i^{th} \) largest eigenvalue of \( S_i \), and \( h_i \) is the \( i^{th} \) largest eigenvalue of \( K_i \).

Numerical Experiments

Data drawn independently from two functions with different GP priors. The change surface between the functions defined by \( \sigma_i(x)^2 \)

\[ \text{where } \sigma_i(x) = \sum_{m=1}^{M} \omega_m \cos(2\pi \langle x \rangle - \langle x' \rangle) \]

We create a predictive test by splitting numerical data into training and testing sets. We compare the GP change surface model to sparse spectrum Gaussian process (SSGP) with 500 basis functions, SS-GP with fixed spectral points with 500 basis functions, and a GP with multiplicative spectral mixture kernels.

Acknowledgments

This material is based upon work supported by the NSF Graduate Research Fellowship under Grant No. DGE 1252522 and the NSF award No. IIS-0953300. Flaxman was supported by EPSRC (EP/K009362/1).