Scalable Gaussian Processes for Characterizing Multidimensional Change Surfaces

William Herlands[1], **Andrew Gordon Wilson**[1], **Hannes Nickisch**[2], **Seth Flaxman**[3], **Daniel Neill**[1], **Wilbert van Panhuis**[4], **Eric Xing**[1]

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

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')) \tag{1}$$
$$m(x) = \mathbb{E}[f(x)] \tag{2}$$
$$k(x, x') = \text{cov}(f(x), f(x')) \tag{3}$$

Any finite collection of function values is normally distributed $[f(x_1)...f(x_n)] \sim \mathcal{N}(m(\boldsymbol{x}), K)$ where $K_{i,j} = k(x_i, x_j)$.

In the case of a Gaussian observation model, $\boldsymbol{y} = f(\boldsymbol{x}) + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon)$, we can express the log marginal likelihood as,

$$\log p(\boldsymbol{y}|\theta) \propto -\log|K + \sigma_\epsilon I| - \boldsymbol{y}^\top (K + \sigma_\epsilon I)^{-1}\boldsymbol{y} \tag{4}$$

GP Change Surface Model

A change surface consists of a convex combination of latent functional regimes, $f_1, \ldots, f_r$.

$$y(x) = \sum_{i=1}^{r} \sigma(w_i(x)) f_i(x) + \epsilon_n \tag{5}$$
$$\sum_{i=1}^{r} \sigma(w_i(x)) = 1 \tag{6}$$

Warping functions $\sigma(z)$

We are particularly interested in latent functions that exhibit some amount of mutual exclusivity. We induce this partial discretization with a warping function, $\sigma(z) : \mathbb{R}^1 \to [0, 1]$, whose range is concentrated towards 0 and 1.

$$\sigma(w_i(x)) = \text{softmax}(\boldsymbol{w}(x))_i = \frac{\exp(w_i(x))}{\sum_{j=1}^{r} \exp(w_j(x))}. \tag{7}$$

Weighting functions $w(x)$

The expressibility of $w(x)$ determines how changes can occur in the data, and how many can occur. 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 Kitchen Sink features.

$$w(x_i) = \sum_{i=1}^{v} a_i \cos(\omega_i^\top x + b_i) \tag{8}$$

Design choices for K

Each latent function is specified by a kernel with unique hyperparameters. In order to maintain maximal generality the model uses spectral mixture kernels where $k_{SM}(\tilde{x}, \tilde{x}') =$

$$\sum_{q=1}^{Q} \omega_q cos(2\pi(\tilde{x} - \tilde{x}')^\top m_q) \prod_{p=1}^{P} \exp(-2\pi^2(\tilde{x}_p - \tilde{x}_p')^2 v_q^{(p)}),$$

where $\tilde{x} \in \mathbb{R}^P$ and $\Sigma_q = \text{diag}(v_q^{(1)}, \ldots, v_q^{(P)})$ is a diagonal covariance matrix for multidimensional inputs.

Nonstationary additive kernel

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

$$k(x, x') = \sum_{i=1}^{r} \sigma(w_i(x)) k_i(x, x') \sigma(w_1 i x')) \tag{9}$$

$\sigma(w_1(x)) \ldots \sigma(w_r(x))$ induce nonstationarity since they are dependent on the input x. Thus, even if we use stationary kernels for all k_i, our model results in a additive, nonstationary kernel.

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-0953330. 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 K^{-1} and $\log(|K|)$. Kronecker methods decompose the covariance matrix $K = K_1 \otimes \cdots \otimes K_D$, where each K_d is $n_d \times n_d$ such that $\prod_1^D n_d = n$, if:

1. inputs lie on a Cartesian grid, $x \in X = X^{(1)} \times ... \times X^{(D)}$
2. kernel is multiplicative across each dimension

Marginal Likelihood	Computation	Memory
Naive	$O(n^3)$	$O(n^2)$
Kronecker	$O(Dn^{\frac{D+1}{D}})$	$O(Dn^{\frac{2}{D}})$

But Kronecker methods are not applicable to additive kernels!

Additive Kronecker approximation

Inverse K^{-1}: use finite difference methods to compute linear conjugate gradients. The key subroutine is MVM so the sum of Kronecker products can be efficiently multiplied by a vector.

Log determinant $\log(|K|)$: Weyl's inequality states that for $n \times n$ Hermitian matrices, $M = A + B$, with sorted eigenvalues $\mu_1, .., \mu_n$, $\alpha_1, ..., \alpha_n$, and $\beta_1, ..., \beta_n$, respectively,

$$\mu_{i+j-1} \leq \alpha_i + \beta_j \tag{10}$$

Thus considering the log determinant,

$$\log(|A + B|) = \log(|M|) = \sum_{i=1}^{n} \log(\mu_i) \tag{11}$$
$$\leq \sum_{i+j-1=1}^{n} \log(\alpha_i + \beta_j) \tag{12}$$

We iteratively apply this approximation to pairs of matrices in order to bound $\log(|\sum_{\ell=1}^{r} K_\ell|)$

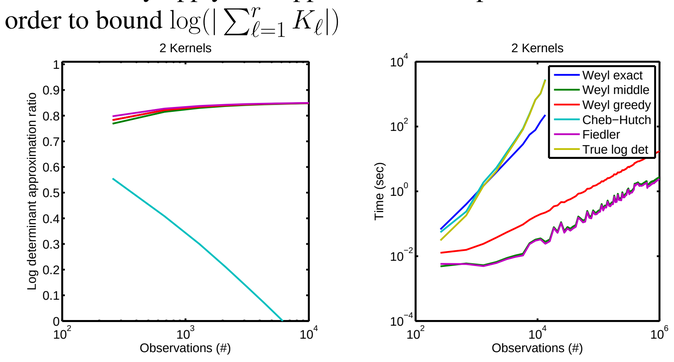

Figure 1: Left shows the approximation ratio to the $\log(|\sum_{i=1}^{2} K_i|)$. Right shows the time to compute each approximation and the truth.

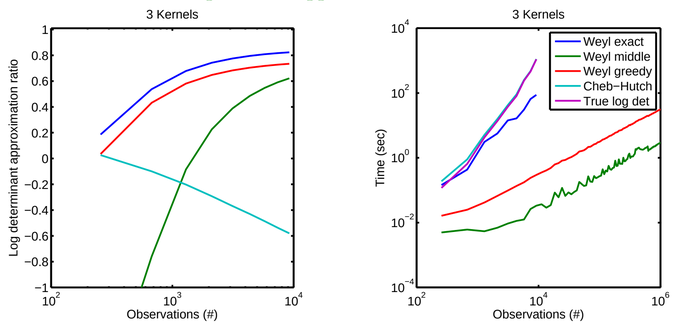

Figure 2: Left shows the approximation ratio to the $\log(|\sum_{i=1}^{3} K_i|)$. Right shows the time to compute each approximation and the truth.

Scaled additive kernels

Rewrite Eq. 9 in matrix notation where $S_i = \text{diag}(\sigma(w_i(x)))$ and $S_i' = \text{diag}(\sigma(w_i(x')))$

$$K = S_1 K_1 S_1' + \cdots + S_r K_r S_r' \tag{13}$$

Employing the bound on eigenvalues of matrix products,

$$\text{sort}(\text{eig}(A * B)) \leq \text{sort}(\text{eig}(A)) * \text{sort}(\text{eig}(B)) \tag{14}$$

we can bound $\log(|K|)$ in Eq. 13 with a Weyl approximation over $[\{s_{i,l} * k_{i,l} * s_{i,l}'\}_{l=1}^{n}]_{i=1}^{r}$ where $s_{i,l}$ is the l^{th} largest eigenvalue of S_i and $k_{i,l}$ is the l^{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 $\sigma(w_{poly}(x))$ where $w_{poly}(x) = \sum_{i=0}^{3} \beta_i^T x^i$, $\beta_i \sim \mathcal{N}(0, I_D)$. We create a predictive test by splitting nuemrical data into training and testing sets. We compare the GP change surface model to sparse spectrum Gaussian process (SSGP) with 500 basis functions, SSGP with fixed spectral points with 500 basis functions, and a GP with multiplicative spectral mixture kernels.

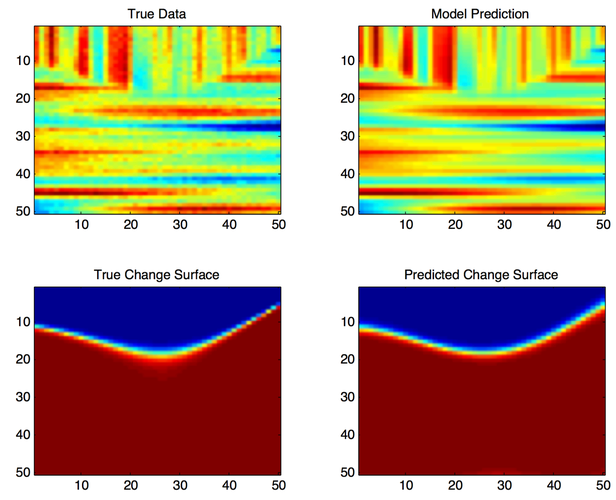

Figure 3: Numerical data experiment. The top-left depicts the data; the bottom-left shows the true change surface, $\sigma(w_1(x))$, blue=0, red=1. The right side depicts the predicted output and change surface.

Method	NMSE
Smooth change surface	0.00078
SSGP	0.01530
SSGP fixed	0.02820
Spectral mixture	0.00200

United States Measles Data

We analyze monthly measles incidence data from 1935 to 2003 in the continental United States. We fit the model to $\approx 33,000$ data points where $x \in \mathbb{R}^3$ with two spatial dimensions representing centroids of each state and one temporal dimension.
Results for three states are shown in Figure 4 along with the predicted change surface. The red line marks the vaccine year of 1963, while the dotted line marks the points where $\sigma(w(x_{state})) = 0.5$. In Figure 5 we depict the midpoint, $\sigma(w(x_{state})) = 0.5$, for each state. In Figure 6 we depict the change surface slope from $\sigma(w(x_{state})) = 0.25$ to $\sigma(w(x_{state})) = 0.75$ for each state to estimate the rate of change.

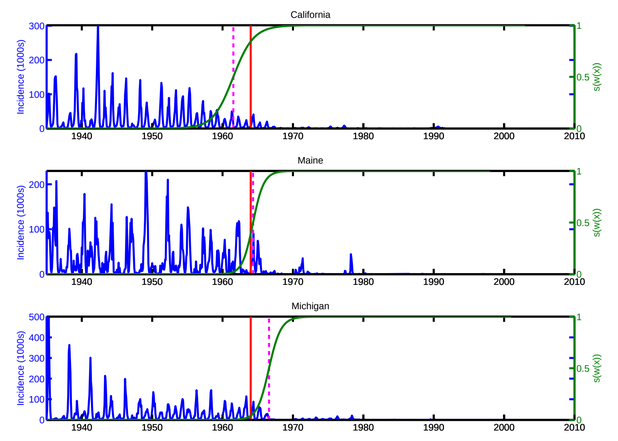

Figure 4: Measles incidence levels from 3 states, 1935 - 2003. The green line plots $\sigma(w(x_{state}))$, the vertical red line indicates the vaccine in 1963, and the magenta line indicates $\sigma(w(x_{state})) = 0.5$.

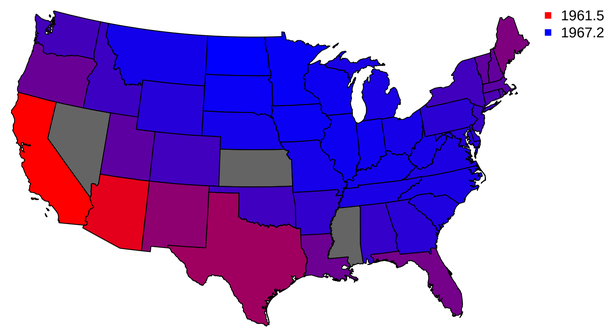

Figure 5: US states colored by the date where $\sigma(w(x_{state})) = 0.5$. Red indicates earlier dates, with California being the earliest. Blue indicates later dates, with North Dakota being the latest. Grayed out states were missing in the dataset.

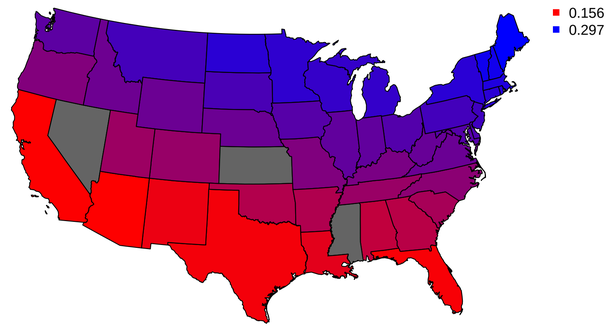

Figure 6: US states colored by the slope of $\sigma(w(x_{state}))$ from 0.25 to 0.75. Red indicates flatter slopes, with Arizona being the lowest. Blue indicates steeper slopes, with Maine being the highest. Grayed out states were missing in the dataset.