Scalable Gaussian Processes for Characterizing Multidimensional Change Surfaces

William Herlands[1] **Andrew 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

Abstract

We present a scalable Gaussian process model for identifying and characterizing smooth multidimensional changepoints, and automatically learning changes in expressive covariance structure. We use Random Kitchen Sink features to flexibly define a *change surface* in combination with expressive spectral mixture kernels to capture the complex statistical structure. Finally, through the use of novel methods for additive non-separable kernels, we can scale the model to large datasets. We demonstrate the model on numerical and real world data, including a large spatio-temporal disease dataset where we identify previously unknown heterogeneous changes in space and time.

1 Introduction

In human systems we are often confronted with changes or perturbations which may not immediately disrupt an entire system. Instead, changes such as policy interventions take time to affect deeply held habits or trickle through a complex bureaucracy. The dynamics of these changes are non-trivial, with sophisticated spatial distributions, rates, and intensity functions. Using expressive models to fully characterize such changes is essential for making accurate predictions and yielding scientifically relevant results.

Typically, changepoint methods (Chernoff and Zacks, 1964) model system perturbations as discrete, or near-discrete, changepoints. These points are either identified sequentially using online algorithms, or retrospectively. Here we consider retrospective analysis (Brodsky and Darkhovsky, 2013; Chen and Gupta, 2011).

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 51.

Gaussian processes have been used for changepoint modeling to provide a nonparametric framework. Saatçi et al. (2010) extend the sequential Bayesian Online Changepoint Detection algorithm (Adams and MacKay, 2007), by using a Gaussian process to model temporal covariance within a particular regime. Similarly, Garnett et al. (2009) provide Gaussian processes for sequential changepoint detection with mutually exclusive regimes. These models focus on discrete changepoints, where regimes defined by distinct Gaussian processes change instantaneously at $t = t_0$. While such models may be appropriate for mechanical systems, they do not permit modeling of the complex changes common to many human systems.

A small collection of pioneering work has briefly considered the possibility of non-discrete Gaussian process change-points (Wilson, 2014; Lloyd et al., 2014). Yet these models rely on sigmoid transformations of linear functions which are restricted to fixed rates of change, and are demonstrated exclusively on small, one-dimensional time series data. They cannot expressively characterize non-linear changes or feasibly operate on large multidimensional data.

Applying changepoints to multiple dimensions, such as spatio-temporal data, is theoretically and practically non-trivial, and has thus been seldom attempted. Notable exceptions include Majumdar et al. (2005) who consider discrete spatio-temporal changepoints with three additive Gaussian processes: one for $t \leq t_0$, one for $t > t_0$, and one for all t. Alternatively, Nicholls and Nunn (2010) use a Bayesian onset-field process on a lattice to model the spatio-temporal distribution of human settlement on the Fiji islands.

The limitations of these models reflect a common criticism that Gaussian processes are unable to convincingly respond to changes in covariance structure. We propose addressing this deficiency with an expressive, flexible, and scalable change surface model.

Throughout the paper we refer to *change surfaces* as the multidimensional generalization of changepoints. Unlike the discrete notion of changepoints, a change

surface can have a variable rate of change and non-monotonicity in the transition between functional regimes. Additionally, changes can occur heterogeneously across the input dimensions. We formalize the notion of a change surface through our model specification in Section 3.

1.1 Main contributions

We introduce a scalable Gaussian process model, which 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 a principled means for scalable kernel learning. Our contributions include:

1. A non-discrete Gaussian process change surface model over multiple input dimensions. Our model specification learns the change surface from data, enabling it to approximate discrete changes or gradual shifts between regimes. The input can have arbitrary dimension, though we primarily focus our attention on spatio-temporal modeling over 2D space and 1D time.

2. The first scalable Gaussian process changepoint model by using novel Kronecker methods. Modern datasets require methods which can scale to hundreds of thousands of instances.

3. A novel method for estimating the log determinant of additive positive semidefinite matrices using the Weyl inequality. This enables scalable additive Gaussian process models with non-separable kernels in space and time.

4. Random Kitchen Sink features to sample from a Gaussian process change surface. This flexibility permits arbitrary changes which can adapt to heterogeneous effects over multiple dimensions. It also permits analytic optimization for the model.

5. We use logistic functions to normalize the weights on all latent functions (one per regime), thereby providing a very interpretable model. Additionally, we permit arbitrary specification of the change surface parameterization, allowing experts to specify interpretable models for how the change surface behaves over the input space.

6. A novel initialization method for spectral mixture kernels by fitting a Gaussian mixture model to the Fourier transform of the data. This provides good starting values for hyperparameters of expressive stationary kernels, allowing for proper optimization over a multimodal parameter space.

7. A nonparametric Bayesian framework for discovering and characterizing continuous changes in large observational data. We demonstrate our approach on numerical and real world data, including a recently developed public health dataset. We demonstrate how the effect of the measles vaccine introduced in the U.S. in 1963 was spatio-temporally varying. Our model discovers the time frame in which the measles vaccine was introduced, and accurately represents the change in dynamics before and after the introduction, thus providing new insights into the spatial and temporal dynamics of reported disease incidence.

1.2 Outline

In the remainder of the paper, Section 2 provides background on Gaussian processes. Section 3 describes our change surface model including the weighting, warping, and kernel functions. Section 4 introduces a novel algorithm for approximating the log determinant of additive kernels. Section 5 details a new initialization procedure for spectral mixture hyperparameters. Section 6 describes our numerical and real-world experiments. Finally, we conclude with summary remarks in section 7.

2 Gaussian Processes

Given data $(\mathbf{y}, \mathbf{x})$, where $\mathbf{y} = \{y_1 ... y_n\}$, are outputs or response variables, and $\mathbf{x} = \{x_1 ... x_n\}, x_i \in \mathbb{R}^D$ are inputs or covariates, we assume that the responses are generated from the inputs by a latent function with a Gaussian process prior and Gaussian noise, such that $\mathbf{y} = f(\mathbf{x}) + \epsilon$, $f(x) \sim GP(m, k)$, $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon)$. A Gaussian process is a nonparametric prior over functions completely specified by mean and covariance functions:

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

Any finite collection of function values is normally distributed $[f(x_1) ... f(x_p)] \sim \mathcal{N}(\boldsymbol{\mu}, K)$ where $\mu_i = m(x_i)$ and $p \times p$ matrix $K_{i,j} = k(x_i, x_j)$.

In order to learn hyperparameters, we often desire to optimize the marginal likelihood of the data, conditioned on kernel hyperparameters θ, and inputs, $\mathbf{x}$.

$$p(\mathbf{y}|\theta, \mathbf{x}) = \int p(y|f, \mathbf{x}) p(f|\theta) df \tag{4}$$

In the case of a Gaussian observation model we can express the log marginal likelihood as,

$$\log p(\mathbf{y}|\theta) \propto -\log |K + \sigma_\epsilon I| - \mathbf{y}^\top (K + \sigma_\epsilon I)^{-1} \mathbf{y} \tag{5}$$

We assume familiarity with the basics of Gaussian processes as described by Rasmussen and Williams (2006).

3 Smooth Change Surface Model

Change surface data consists of latent functions $f_1, \ldots, f_r$ defining r regimes in the data. The transition between any two functions is considered a change surface. Were these r functions not mutually exclusive, we could consider an input dependent mixture model such as (Wilson et al., 2012),

$$y(x) = w_1(x)f_1(x) + \cdots + w_r(x)f_r(x) + \epsilon_n \tag{6}$$

where the weighting functions, $w_i(x) : \mathbb{R}^D \to \mathbb{R}^1$, describe the mixing proportions over the input domain. However, for data with changing regimes we are particularly interested in latent functions that exhibit some amount of mutual exclusivity.

We induce this partial discretization with a warping function, $\sigma(z) : \mathbb{R}^1 \to [0, 1]$, which has support over the entire real line but a range which is concentrated towards 0 and 1. Additionally, we choose $\sigma(z)$ such that it produces a convex combination over the weighting functions, $\sum_{i=1}^{r} \sigma(w_i(x)) = 1$. In this way, each $w_i(x)$ defines the strength of latent f_i over the domain, while $\sigma(z)$ normalizes these weights to induce weak mutual exclusivity.

A natural choice for flexible, smooth change surfaces is the softmax function since it can approximate a Heaviside step function or gradual changes. For r latent functions, the resulting warping function is

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

Our model is thus,

$$y(x) = \sigma(w_1(x))f_1(x) + \cdots + \sigma(w_r(x))f_r(x) + \epsilon_n \tag{8}$$

If we assume Gaussian process priors on all latent functions $f_1(x), \ldots, f_r(x)$ we can define $y(x) = f(x) + \epsilon$ where $f(x)$ has a Gaussian process prior with covariance function,

$$\begin{aligned} k(x, x') = \sigma(w_1(x))k_1(x, x')\sigma(w_1(x'))+ \\ \cdots + \sigma(w_r(x))k_r(x, x')\sigma(w_r(x')) \end{aligned} \tag{9}$$

This assumption does not limit the expressiveness of Eq. 8 since each Gaussian process may be defined with different mean and covariance functions. Indeed, where the data exhibits latent functional change we expect that the latent functions will have correspondingly different hyperparameters even if the kernel forms are identical.

$\sigma(w_1(x)) \ldots \sigma(w_r(x))$ 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 flexible, nonstationary kernel.

Each $\sigma(w_i(x))$ defines how the coverage of $f_i(x)$ varies over the input domain. Where $\sigma(w_i(x)) \approx 1$, $f_i(x)$ dominates and primarily describes the relationship between $\mathbf{x}$ and $\mathbf{y}$, and in cases where there is no i such that $\sigma(w_i(x)) \approx 1$, a number of functions are dominant in defining the relationship between $\mathbf{x}$ and $\mathbf{y}$. Since $\sigma(z)$ pushes values towards 1 or 0, the regions with multiple dominant functions are transitory and thus considered change regions. Therefore, we can interpret how the change surface develops and where different regimes dominate by evaluating $\sigma(w(x))$ over the input domain.

3.1 Design choices for $w(x)$

The functional form of $w(x)$ determines how changes can occur in the data, and how many can occur. For example, a linear parametric weighting function,

$$w(x) = \beta_0 + \beta_1^\top x\,, \tag{10}$$

only permits a single linear change surface in the data. Yet even this simple model is more expressive than discrete changepoints since it permits flexibility in the rate of change and extends to change regions in $\mathbb{R}^D$.

In order to develop a general framework we do not require any prior knowledge about the functional form of $w(x)$ and instead assume a Gaussian process prior on $w(x)$. While in principle we could sample from the full Gaussian process prior, this would lead to a non-conjugate model which would thus be less computationally attractive and significantly constrain the "plug and play" nature of choices for $\sigma(z)$, $w(x)$, and K. Instead, we approximate the Gaussian process with Random Kitchen Sink (RKS) features and analytically derive inference procedures using the log marginal likelihood (Lázaro-Gredilla et al., 2010).

Rahimi and Recht (2007) demonstrate that if we consider the vector of RKS features, $\phi(x) : \mathbb{R}^D \to \mathbb{R}^m$ with $\omega_i \in \mathbb{R}^D$,

$$\phi(x)^\top = \sqrt{\frac{2}{m}}[\cos(\omega_i^\top x + b_i)]_{i=1}^m \tag{11}$$

then we can approximate any stationary kernel by taking the Fourier transform of $k(x - x') = k(\delta)$,

$$p(\omega) = \frac{1}{2\pi}\int \exp(-j\omega\delta)k(\delta)d\delta \tag{12}$$

and putting priors over the parameters of the RKS feature mapping,

$$\omega_i \sim p(\omega) \tag{13}$$

$$b_i \sim \text{Uniform}(0, 2\pi) \tag{14}$$

For an RBF kernel where $\Lambda = \text{diag}(l_1^2, \ldots, l_D^2)$ is a diagonal matrix of length-scales, we sample,

$$\omega_i \sim \mathcal{N}(0, \frac{1}{4\pi^2}\Lambda^{-1}) \tag{15}$$

Therefore, if we want to place a Gaussian process prior over our weighting functions, $w(x) \sim GP(0, K)$, we can use RKS features to create a compact representation of the kernel (Lázaro-Gredilla et al., 2010). For any finite input $\mathbf{x}$ we know that,

$$g(\mathbf{x}) \sim \mathcal{N}(0, K) \tag{16}$$

Equivalently, we can define parameters a such that,

$$a \sim \mathcal{N}(0, \frac{\sigma_0}{m} I) \tag{17}$$

$$w(\mathbf{x}) = \phi(\mathbf{x})^\top a \tag{18}$$

which we can write in the explicit RKS feature space representation,

$$w(x_i) = \sum_{i=1}^{v} a_i \cos(\omega_i^\top x + b_i) \tag{19}$$

allowing us to sample from $w(x)$ with a finite sum of RKS features. Initialization of the hyperparameters σ_0 and Λ is discussed in the supplementary material.

Experts with domain knowledge can specify a parametric form for $w(x)$ other than RKS features. Such specification can be advantageous, requiring relatively few, highly interpretable parameters to optimize.

3.2 Design choices for K

Each latent function is specified by a kernel with unique hyperparameters. By design, each k_i may be of a different form. For example, one function may have a Matérn kernel, another a periodic kernel, and a third an exponential kernel. Such specification is useful when domain knowledge provides insight into the covariance structure of the various regimes.

In order to maintain maximal generality and expressivity, we develop the model using spectral mixture kernels (Wilson and Adams, 2013) 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. With a sufficiently large Q, spectral mixture kernels can approximate any stationary kernel, providing the flexibility to capture complex patterns over multiple dimensions. These kernels have been used in pattern prediction, outperforming complex combinations of standard stationary kernels (Wilson et al., 2014).

Using spectral mixture kernels extends previous work on Gaussian processes changepoint modeling which has been restricted in practice to RBF (Saatçi et al., 2010; Garnett et al., 2009) or exponential kernels (Majumdar et al., 2005). Expressive covariance functions are particularly important with multidimensional and spatio-temporal data where the dynamics are complex and unknown a priori. While most Gaussian process models provide the theoretical flexibility to choose any kernel, the practical mechanics of initializing and fitting more expressive kernels is a challenging problem. We describe an initialization procedure in Section 5 which we hope can enable other models to exploit expressive kernels as well.

4 Scalable inference

Analytic optimization and inference requires computation of the log marginal likelihood (Eq. 5). Yet calculating the inverse and log determinant of $n \times n$ covariance matrices requires $O(n^3)$ computations and $O(n^2)$ memory (Rasmussen and Williams, 2006), which is impractical for large datasets. Recent advances in scalable Gaussian processes have reduced this computational burden by exploiting Kronecker structure under two assumptions. One, the inputs lie on a grid formed by a Cartesian product, $x \in X = X^{(1)} \times ... \times X^{(D)}$. Two, the kernel is multiplicative across each dimension. The assumption of separable, multiplicative kernels is commonly employed in spatio-temporal Gaussian process modeling (Martin, 1990; Majumdar et al., 2005; Flaxman et al., 2015). Under these assumptions, the $n \times n$ covariance matrix $K = K_1 \otimes \cdots \otimes K_D$, where each K_d is $n_d \times n_d$ such that $\prod_1^D n_d = n$.

Using efficient Kronecker algebra, Saatçi (2012) calculates the inverse and log determinant calculations in $O(Dn^{\frac{D+1}{D}})$ operations using $O(Dn^{\frac{2}{D}})$ memory. Furthermore, Wilson et al. (2014) extends the Kronecker methods for incomplete grids.

Yet for an additive kernel such as that needed for change surface modeling (Eq. 9), calculating the inverse and log determinant is no longer feasible using Kronecker algebra as in Saatçi (2012) because the sum of the matrix Kronecker products does not decompose as a single Kronecker product. Instead, calculations involving the inverse can be efficiently carried out using finite difference methods to compute linear conjugate gradients as in Flaxman et al. (2015) because the key subroutine is matrix-vector multiplication and the sum of Kronecker products can be efficiently multiplied by a vector.

However, there is no exact method for efficient computation of the log determinant of the sum of Kronecker products. Instead, Flaxman et al. (2015) upper bound the log determinant using the Fiedler bound (Fiedler, 1971) which says that for $n \times n$ Hermitian matrices A and B with sorted eigenvalues $\alpha_1, \ldots, \alpha_n$

and $\beta_1, \ldots, \beta_n$ respectively,

$$\log(|A+B|) \leq \sum_{i=1}^{n} \log(\alpha_i + \beta_{n-i+1}) \tag{20}$$

While this yields fast, $O(n)$ computation, the Fiedler bound does not generalize for more than two matrices.

Instead, we bound the log determinant of the sum of multiple covariance matrices using Weyl's inequality (Weyl, 1912) which states that for $n \times n$ Hermitian matrices, $M = A + B$, with sorted eigenvalues $\mu_1, \ldots, \mu_n$, $\alpha_1, \ldots, \alpha_n$, and $\beta_1, \ldots, \beta_n$, respectively,

$$\mu_{i+j-1} \leq \alpha_i + \beta_j \tag{21}$$

Since $\log(|A+B|) = \log(|M|) = \sum_{i=1}^{n} \log(\mu_i)$ we can bound the log determinant by $\sum_{i+j-1=1}^{n} \log(\alpha_i + \beta_j)$. Furthermore, we can use the Weyl bound iteratively over pairs of matrices to bound the sum of r covariance matrices $K_1, \ldots, K_r$.

As the bound indicates, there is flexibility in the choice of which eigenvalue pair $\{\alpha_i, \beta_j\}$ to sum in order to bound μ_{i+j-1}. One might be tempted to minimize over all possible pairs for each of the n eigenvalues of M in order to obtain the tightest bound on the log determinant. Unfortunately, this requires $O(n^2)$ computations. Instead we explore two possible alternatives:

1. For each μ_{i+j-1} we choose the "middle" pair such that $i = j$ when possible, and $i = j+1$ otherwise. This heuristic requires $O(n)$ computations.

2. We employ a greedy search by using the previous i' and j' to choose the minimum of $2s$ pairs of eigenvalues $\{\alpha_i, \beta_j\}_{i=i'-s}^{i=i'+s}$. When $s = 0$ this corresponds to the middle heuristic. When $s = \frac{n}{2}$ this corresponds to the exact Weyl bound. The greedy search requires $O(2sn)$ computations.

In addition to bounding the sum of kernels, we must also deal with the scaling functions, $\sigma(w_i(x))$. We can rewrite Eq. 9 in matrix notation,

$$K = S_1 K_1 S_1' + \cdots + S_r K_r S_r' \tag{22}$$

where $S_i = \text{diag}(\sigma(w_i(x)))$ and $S_i' = \text{diag}(\sigma(w_i(x')))$. Employing the bound on eigenvalues of matrix products (Bhatia, 2013),

$$\text{sort}(\text{eig}(A * B)) \leq \text{sort}(\text{eig}(A)) * \text{sort}(\text{eig}(B)) \tag{23}$$

we can bound the log determinant of K in Eq. 22 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

We empirically evaluate the exact Weyl bound, middle heuristic, and greedy search with $s = 40$ for our model using synthetic data (generated according to the procedure in Section 6.1). We compare these results against the Fiedler bound (in the case of two kernels), and a recently proposed method for estimating the log determinant using Chebyshev polynomials coupled with stochastic Hutchinson trace approximation (Han et al., 2015). Figures 1 and 2 depict the ratio of each approximation to the true log determinant, and the time to compute each approximation over increasing number of observations for 2 and 3 kernels. We note that all Weyl and Fiedler approximations converge to ≈ 0.8 of the true log determinant, which was negative in the experiments. While the exact Weyl bound scales poorly, as expected, both approximate Weyl bounds scale well. In practice, we use the middle heuristic since it provides the fastest results.

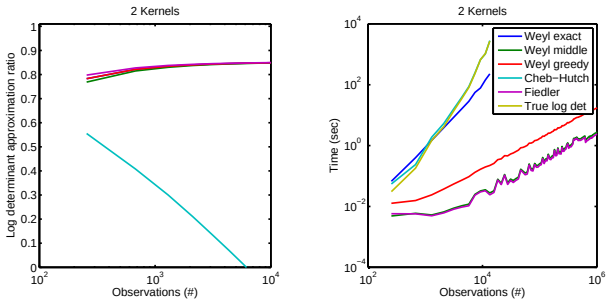

Figure 1: Left shows the approximation ratio to the log determinant of 2 additive kernels. Right shows the time to compute each approximation and the truth.

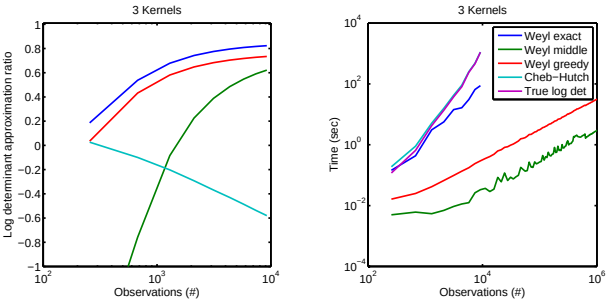

Figure 2: Left shows the approximation ratio to the log determinant of 3 additive kernels. Right shows the time to compute each approximation and the truth.

5 Initialization

Since our model uses expressive spectral mixture kernels and flexible RKS features, the parameter space is highly multimodal. Therefore, it is essential to initialize the model hyperparameters appropriately. The supplementary material includes an initialization algorithm for the $w(x)$ RKS hyperparameters.

Assuming an initialized $w(x)$ we define the subset $\{x :$

$\sigma(w_i(x)) > 0.5\}$ where each latent function f_i from Eq. 8 is dominant. We then take a Fourier transform of $y(x)$ over each dimension, $x^{(d)}$, of $\{x : \sigma(w_i(x)) > 0.5\}$ to obtain the empirical spectrum in that dimension. Note that we consider each dimension of x individually since we have a multiplicative Q-component spectral mixture kernel over each dimension. Since spectral mixture kernels model the spectral density with Q Gaussians on $\mathbb{R}^1$, we fit a 1D Gaussian mixture model,

$$p(x) = \sum_{q=1}^{Q} \phi_q \mathcal{N}(\mu_q, \sigma_q) \tag{24}$$

to the the empirical spectrum for each dimension. Using the learned mixture model we initialize the parameters of our spectral mixture kernels for $f_i(x)$.

Algorithm 1 Initialize spectral mixture kernels

1: **for** $k_i : i = 1 : r$ **do**
2: **for** $d = 1 : D$ **do**
3: Compute $x^{(d)} \in \{x : \sigma(w_i(x)) > 0.5\}$
4: Sample $s \sim |FFT(sort(y(x^{(d)})))|^2$
5: Fit Q component 1D GMM to s
6: Initialize $\omega_q = std(y) * \phi_q$; $m_q = \mu_q$; $v_q = \sigma_q$
7: **end for**
8: **end for**

After initializing $w(x)$ and spectral mixture hyperparameters, we jointly optimize the entire model using marginal likelihood and standard gradient techniques (Rasmussen and Nickisch, 2010).

6 Experiments

We test our model with both numerical and real world data. There do not exist standard datasets for evaluating spatio-temporal changepoint models. For example, Majumdar et al. (2005) used simulations to demonstrate the effectiveness of their model. Therefore, we apply our method on a standard 1D changepoint dataset, synthetic data, and a newly available spatio-temporal disease dataset.

6.1 Numerical Experiments

We generate a 50×50 grid of synthetic data by drawing independently from two latent functions. Each function is characterized by a 2D RBF kernel with different length-scales and variances. The synthetic change surface between the functions is defined by $\sigma(w_{poly}(x))$ where $w_{poly}(x) = \sum_{i=0}^{3} \beta_i^T x^i$, $\beta_i \sim \mathcal{N}(0, 3I_D)$.

We apply our change surface model with two latent functions, spectral mixture kernels, and $w(x)$ defined by 5 RKS features. We do not provide the model prior information about the change surface or latent functions. Figures 3 and 4 depict typical results using the initialization procedure followed by analytic optimization. The model captures the change surface and produces an appropriate regression over the data.

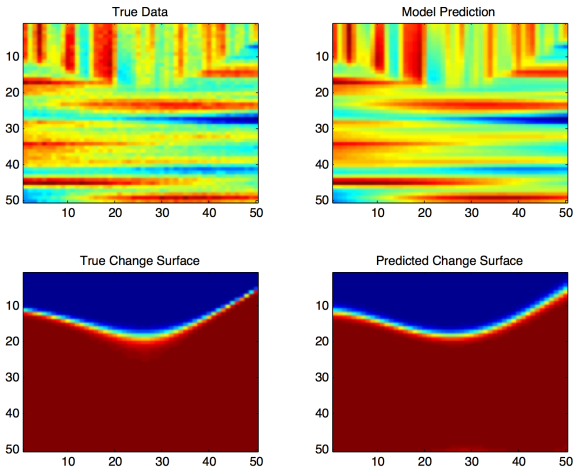

Figure 3: Numerical data experiment. The top-left depicts the data; the bottom-left shows the true change surface with the range from blue to red depicting $\sigma(w_1(x))$. The top-right depicts the predicted output; the bottom-right shows the predicted change surface.

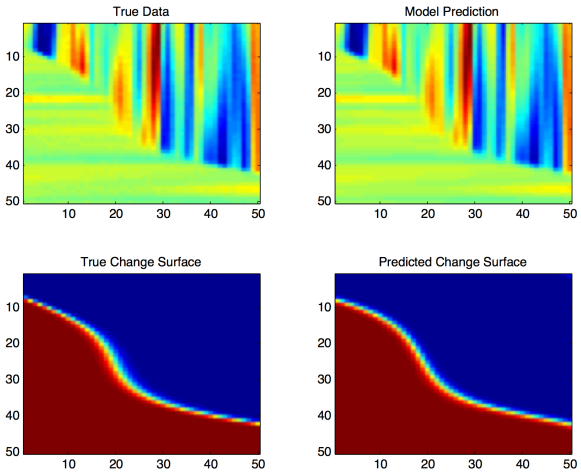

Figure 4: Numerical data experiment. The top-left depicts the data; the bottom-left shows the true change surface with the range from blue to red depicting $\sigma(w_1(x))$. The top-right depicts the predicted output; the bottom-right shows the predicted change surface.

Using synthetic data, we create a predictive test by splitting the data into training and testing sets. We compare our smooth change surface model to three other expressive, scalable methods: sparse spectrum Gaussian process with 500 basis functions (Lázaro-Gredilla et al., 2010), sparse spectrum Gaussian process with fixed spectral points with 500 basis functions (Lázaro-Gredilla et al., 2010), and a Gaussian process

Table 1: Comparison of prediction using flexible, scalable Gaussian process methods on synthetic multidimensional change-surface data.

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

with multiplicative spectral mixture kernels in each dimension. For each method we average the results for 10 random restarts. Table 1 shows the normalized mean squared error (NMSE) of each method,

$$\text{NMSE} = \frac{\|y_{test} - y_{pred}\|_2^2}{\|y_{test} - \bar{y}_{train}\|_2^2} \tag{25}$$

where $\bar{y}_{train}$ is the mean of the training data.

Our change surface model performed best due to the expressive nonstationary covariance function that fits to the different functional regimes in the data. Although the alternate methods can flexibly adapt to the data, they must account for the change in covariance structure by setting an effectively shorter length-scale over the data. Thus their predictive accuracy is reduced compared to the change surface model.

6.2 British Coal Mining Data

British coal mining accidents from 1861 to 1962 have been well studied in the point process and changepoint literature (Raftery and Akman, 1986; Adams and MacKay, 2007). We use yearly counts of accidents from Carlin et al. (1992). Domain knowledge suggests that the Coal Mines Regulation Act of 1887 affected the underlying process of coal mine accidents. This act limited child labor in mines, detailed inspection procedures, and regulated construction standards.

We apply our change surface model with two latent functions, spectral mixture kernels, and $w(x)$ defined by 5 RKS features. We do not provide the model with prior information about the 1887 legislation date. Figure 5 depicts the cumulative data and predicted change surface. The red line marks the year 1887 and the magenta line marks $x : \sigma(w(x)) = 0.5$. Our algorithm correctly identified the change region and suggests a gradual change that took 5.6 years to transition from $\sigma(w_1(x)) = 0.25$ to $\sigma(w_1(x)) = 0.75$.

Using the coal mining data we apply a number of well known univariate changepoint methods using their standard settings. We compared Pruned Exact Linear Time (PELT) (Killick et al., 2012) for changes in mean and variance and a nonparametric method named "ecp" (James and Matteson, 2013). Additionally, we tested the batch changepoint method described in Ross (2013) with Student-t and Bartlett tests for Gaussian data as well as Mann-Whitney and Kolmogorov-Smirnov tests for nonparametric changepoint estimation. Figure 2 compares the dates of change identified by these methods to the date where $\sigma(w_1(x)) = 0.5$ in our method.

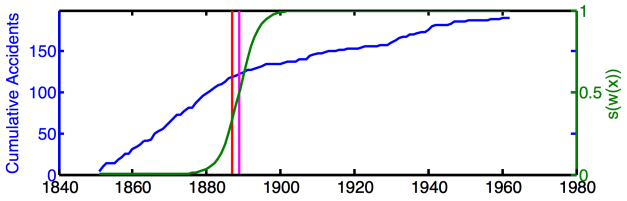

Figure 5: British coal mining accidents from 1851 to 1962. The blue line depicts cumulative annual accidents, the green line plots $\sigma(w(x))$, the vertical red line marks the Coal Mines Regulation Act of 1887, and the vertical magenta line indicates $\sigma(w_1(x)) = 0.5$.

Table 2: Comparing methods for estimating the date of change in coal mining data.

Method	Estimated date
Change surface $\sigma(w_1(x)) = 0.5$	1888.8
PELT mean change	1886.5
PELT variance change	1882.5
ecp	1887
Student-t test	1886.5
Bartlett test	1947.5
Mann-Whitney test	1891.5
Kolmogorov-Smirnov test	1896.5

Most of the methods identified a change date between 1886 and 1895 except the Bartlett test. While each method provides a point estimate of the change date, only the the change surface model yields a clear analysis of the development of this change. Indeed the 5.6 years that the change surface transitions between $\sigma(w_1(x)) = 0.25$ to $\sigma(w_1(x)) = 0.75$ well encapsulates most of the point estimate method results.

6.3 United States Measles Data

Measles was nearly eradicated in the United States following the introduction of the measles vaccine in 1963. We analyze monthly incidence data for measles from 1935 to 2003 in each of the continental United States and the District of Columbia, made publicly available by Project Tycho (van Panhuis et al., 2013). 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.

We apply our change surface model with two latent

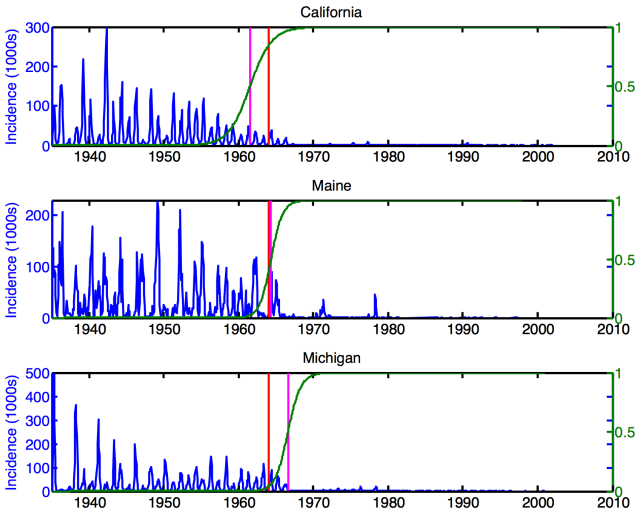

Figure 6: Measles incidence levels from 3 states, 1935 - 2003. The green line plots $\sigma(w(x_{state}))$, the vertical red line indicates the vaccine in 1963, and the magenta line indicates $\sigma(w(x_{state})) = 0.5$.

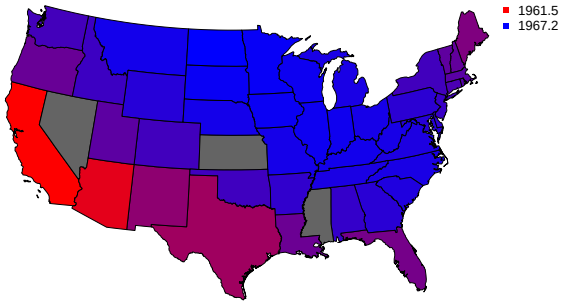

Figure 7: US states colored by the date where $\sigma(w(x_{state})) = 0.5$. Red indicates earlier dates, with California being the earliest. Blue indicates later dates, with North Dakota being the latest. Grayed out states were missing in the dataset.

functions, spectral mixture kernels, and $w(x)$ defined by 5 RKS features. We do not provide prior information about the 1963 vaccination date.

Results for three states are shown in Figure 6 along with the predicted change surface. The red line marks the vaccine year of 1963, while the magenta line marks the points where $\sigma(w(x_{state})) = 0.5$. Our algorithm correctly identified the time frame when the measles vaccine was released in the US.

Additionally, the model suggests that the effect of the measles vaccine varied both temporally and spatially. In Figure 7 we depict the midpoint, $\sigma(w(x_{state})) = 0.5$, for each state. We discover that there is an approximately 6 year difference in midpoint between states. In Figure 8 we depict the change surface slope from

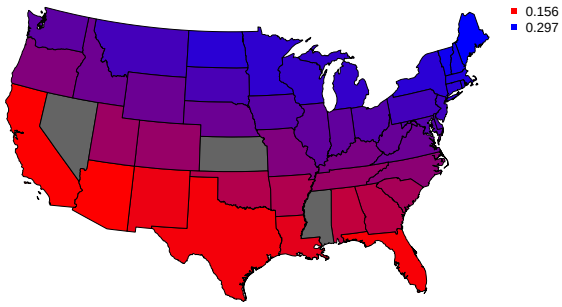

Figure 8: US states colored by the slope of $\sigma(w(x_{state}))$ from 0.25 to 0.75. Red indicates flatter slopes, with Arizona being the lowest. Blue indicates steeper slopes, with Maine being the highest. Grayed out states were missing in the dataset.

$\sigma(w(x_{state})) = 0.25$ to $\sigma(w(x_{state})) = 0.75$ for each state to estimate the rate of change. Here we find that some states had approximately twice the rate of change as others. These variations in the change surface illustrate how the measles vaccine affected states heterogeneously over space and time. They suggest that further scientific research is warranted to understand the underlying causes of this heterogeneity in order to provide insight for future vaccination programs.

7 Conclusions

We presented a scalable, multidimensional Gaussian process model with expressive kernel structure which can learn a complex change surface from data. Using the Weyl inequality, we perform efficient inference with additive kernel structure using Kronecker methods, enabling a multidimensional non-separable kernel. Additionally, we introduce a novel initialization algorithm for learning the $w(x)$ RKS features and spectral mixture kernels. Finally, we apply our model to numerical and real world data, illustrating how it can characterize heterogeneous spatio-temporal change surfaces, yielding scientifically relevant insights.

The work on changepoint modeling is extensive and the current work cannot address all facets of the literature. Future work can extend our retrospective analysis to address sequential change surface detection. Additionally, the method can be extended to automatically determine the number of latent functions.

Acknowledgements

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)

References

Adams, R. P. and MacKay, D. J. (2007). Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*.

Bhatia, R. (2013). *Matrix analysis*, volume 169. Springer Science & Business Media.

Brodsky, E. and Darkhovsky, B. S. (2013). *Nonparametric methods in change point problems*, volume 243. Springer Science & Business Media.

Carlin, B. P., Gelfand, A. E., and Smith, A. F. (1992). Hierarchical bayesian analysis of changepoint problems. *Applied statistics*, pages 389–405.

Chen, J. and Gupta, A. K. (2011). *Parametric statistical change point analysis: With applications to genetics, medicine, and finance*. Springer Science & Business Media.

Chernoff, H. and Zacks, S. (1964). Estimating the current mean of a normal distribution which is subjected to changes in time. *The Annals of Mathematical Statistics*, pages 999–1018.

Fiedler, M. (1971). Bounds for the determinant of the sum of hermitian matrices. *Proceedings of the American Mathematical Society*, pages 27–31.

Flaxman, S. R., Wilson, A. G., Neill, D. B., Nickisch, H., and Smola, A. J. (2015). Fast kronecker inference in gaussian processes with non-gaussian likelihoods. *International Conference on Machine Learning 2015*.

Garnett, R., Osborne, M. A., and Roberts, S. J. (2009). Sequential bayesian prediction in the presence of changepoints. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 345–352. ACM.

Han, I., Malioutov, D., and Shin, J. (2015). Large-scale log-determinant computation through stochastic chebyshev expansions. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 908–917.

James, N. A. and Matteson, D. S. (2013). ecp: An r package for nonparametric multiple change point analysis of multivariate data. *arXiv preprint arXiv:1309.3295*.

Killick, R., Fearnhead, P., and Eckley, I. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598.

Lázaro-Gredilla, M., Quiñonero-Candela, J., Rasmussen, C. E., and Figueiras-Vidal, A. R. (2010). Sparse spectrum gaussian process regression. *The Journal of Machine Learning Research*, 11:1865–1881.

Lloyd, J. R., Duvenaud, D., Grosse, R., Tenenbaum, J., and Ghahramani, Z. (2014). Automatic construction and natural-language description of nonparametric regression models. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Majumdar, A., Gelfand, A. E., and Banerjee, S. (2005). Spatio-temporal change-point modeling. *Journal of Statistical Planning and Inference*, 130(1):149–166.

Martin, R. (1990). The use of time-series models and methods in the analysis of agricultural field trials. *Communications in Statistics-Theory and Methods*, 19(1):55–81.

Nicholls, G. K. and Nunn, P. D. (2010). On building and fitting a spatio-temporal change-point model for settlement and growth at bourewa, fiji islands. *arXiv preprint arXiv:1006.5575*.

Raftery, A. and Akman, V. (1986). Bayesian analysis of a poisson process with a change-point. *Biometrika*, pages 85–89.

Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184.

Rasmussen, C. E. and Nickisch, H. (2010). Gaussian processes for machine learning (gpml) toolbox. *The Journal of Machine Learning Research*, 11:3011–3015.

Rasmussen, C. E. and Williams, C. K. (2006). Gaussian processes for machine learning.

Ross, G. J. (2013). Parametric and nonparametric sequential change detection in r: The cpm package. *Journal of Statistical Software*, page 78.

Saatçi, Y. (2012). *Scalable inference for structured Gaussian process models*. PhD thesis, University of Cambridge.

Saatçi, Y., Turner, R. D., and Rasmussen, C. E. (2010). Gaussian process change point models. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 927–934.

van Panhuis, W. G., Grefenstette, J., Jung, S. Y., Chok, N. S., Cross, A., Eng, H., Lee, B. Y., Zadorozhny, V., Brown, S., Cummings, D., et al. (2013). Contagious diseases in the united states from 1888 to the present. *The New England journal of medicine*, 369(22):2152.

Weyl, H. (1912). Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479.

Wilson, A. and Adams, R. (2013). Gaussian process kernels for pattern discovery and extrapolation. In *Proceedings of The 30th International Conference on Machine Learning*, pages 1067–1075.

Wilson, A., Ghahramani, Z., and Knowles, D. A. (2012). Gaussian process regression networks. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 599–606.

Wilson, A., Gilboa, E., Cunningham, J. P., and Nehorai, A. (2014). Fast kernel learning for multidimensional pattern extrapolation. In *Advances in Neural Information Processing Systems*, pages 3626–3634.

Wilson, A. G. (2014). *Covariance kernels for fast automatic pattern discovery and extrapolation with Gaussian processes*. PhD thesis, PhD thesis, University of Cambridge.

Supplementary Material
Scalable Gaussian Processes for Characterizing Multidimensional Change Surfaces

William Herlands[1] **Andrew 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

Initialization of $w(x)$ RKS Features

To initialize $w(x)$ defined by RKS features we first simplify our change surface model and assume that each latent function $f_1, ..., f_r$ from Eq. 8 is drawn from a Gaussian process with an RBF kernel. Since RBF kernels have many fewer hyperparameters than spectral mixture kernels, this enables the initialization to focus on $w(x)$. Algorithm 2 provides the procedure for initializing this simplified change surface model. Note that depending on the application domain, a model with latent functions defined by RBF kernels may be sufficient.

Algorithm 2 Initialize RKS $w(x)$ by optimizing a simplified model with RBF kernels

1: **for** $i = 1 : g$ **do**
2: Draw a, ω, b for RKS features in $w(x)$
3: Draw h random values for RBF kernels. Choose the best with maximum marginal likelihood
4: Partial optimization of $w(x)$ and RBF kernels
5: **end for**
6: Choose the best set of hyperparameters with maximum marginal likelihood
7: Optimize all hyperparameters until convergence

In the algorithm, we test multiple possible sets of values for $w(x)$ by drawing the hyperparameters a, ω, and b from their respective prior distributions g number of times. To recall the prior distributions from Section 3.1 were,

$$a \sim \mathcal{N}(0, \frac{\sigma_0}{m} I) \tag{1}$$

$$\omega_i \sim \mathcal{N}(0, \frac{1}{4\pi^2} \Lambda^{-1}) \tag{2}$$

$$b_i \sim \text{Uniform}(0, 2\pi) \tag{3}$$

We set reasonable values for hyperparameters in the prior distributions. Specifically, we let $\Lambda = (\frac{range(x)}{2})^2$, $\sigma_0 = \text{std}(y)$, and $\sigma_n = \frac{\text{mean}(|y|)}{10}$. These choices are similar to those used in Lázaro-Gredilla et al. (2010).

For each set of $w(x)$ hyperparameters that we sample, we sample sets of hyperparameters for the RBF kernels h number of times and select the set that yeilds the maximum marginal likelihood. Then we run an abbreviated optimization procedure over each set of $w(x)$ and RBF hyperparameters and finally select the joint set that yeilds the maximum marginal likelihood. Finally, we optimize all the resulting parameters until convergence.

References

Lázaro-Gredilla, M., Quiñonero-Candela, J., Rasmussen, C. E., and Figueiras-Vidal, A. R. (2010). Sparse spectrum gaussian process regression. *The Journal of Machine Learning Research*, 11:1865–1881.