Scalable Gaussian Processes for Characterizing Multidimensional Change Surfaces

April 18, 2016

William Herlands[1]
Carnegie Mellon University
Machine Learning and Public Policy

Advisors: Daniel Neill [2], Alex Smola[3], Wilbert Van Panhuis[4]

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. Through the use of novel methods for additive non-separable kernels, we scale the model to large datasets. We demonstrate the model on numerical simulations as well as applying it to real world spatio-temporal data. Specifically, we model state level incidence rates of measles in the United States both before and after the introduction of the measles vaccine. Additionally we model zip code level requests for lead testing kits in New York City over the past two years in the midst of heightened concerns about lead-tainted water.

[1]Email: herlands@cmu.edu
[2]Email: neill@cs.cmu.edu
[3]Email: alex@smola.org
[4]Email: WAV10@pitt.edu

Contents

1 Introduction **3**
1.1 Literature review 4
1.2 Main contributions 5
1.3 Outline 7

2 Gaussian Processes **7**

3 Methodology **8**
3.1 Smooth Change Surface Model 8
3.1.1 Design choices for $w(x)$ 9
3.1.2 Design choices for K 11
3.2 Scalable inference 12
3.3 Initialization 14

4 Results **17**
4.1 Numerical Experiments 17
4.2 British Coal Mining Data 18
4.3 United States Measles Data 20
4.4 New York City Lead Data 26

5 Conclusion **31**

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 and natural disasters take time to affect deeply engrained habits or trickle through a complex bureaucracy. The dynamics of these changes are non-trivial, with sophisticated distributions, rates, and intensity functions. Specifically, in the spatio-temporal domain changes are often heterogeneously distributed across space and time. Capturing the complexity of these changes can provide useful insight for future policy makers or scientists, enabling them to better target interventions, structure policy interventions, or predict natural phenomena.

To concretize the notion of complex changes, 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.

We develop a highly expressive Gaussian process model able to characterize complex changes and model multiple functional regimes simultaneously. By exploring the potential demographic, political, and natural factors that affect the contours of the change surface, this model can provide insight to policy makers.

In Section 4 we use our model to analyze two real world spatio-temporal datasets and provide policy and scientifically relevant insights. First, we analyze monthly incidence data for measles from 1935 to 2003 in each of the continental United States and the District of Columbia. Our model identifies a change between two distinct regimes in the 1960s. The change surface varies heterogeneously over space and time, with the midpoint of the change surface in each state occurring between 1961 and 1968. This corresponds to the introduction of the measles vaccine in 1963. The heterogeneity that we observed is important since it may indicate differential penetration or effectiveness of the measles vaccine program in the initial years of its implementation. We conduct a preliminary regression analysis to understand the sources of this heterogeneity and discover that later change dates are significantly correlated with family income inequality. This suggests that delayed effectiveness of the measles vaccination program may have been due to difficulties with implementation in rural and economically disadvantaged communities. Additionally, our results show that

states with later change dates had steeper change surfaces and thus switched regimes more quickly. This may have been a result of increased immunity nationwide making it easier for states with later change dates to effectively control the disease in their borders. These conclusions can provide insight to policy makers interested in structuring future vaccination programs, encouraging them to focus on rural communities and ensure reliable delivery to socioeconomically disadvantaged areas.

We also analyze a dataset related to concerns about lead-tainted water in New York City. Since 2015, concerns about lead poisoning in Flint, Michigan's water supply have garnered national attention. We used weekly requests for residential lead testing kits in New York City between January 2014 and April 2016 as a proxy for measuring the concern about lead tainted water in each zip code of the city. Unlike the measles data, there is no ground truth change or single event of importance in this domain. Applying our change surface model we identify a change surface with shifts occurring mostly in 2015 and with distinct geographic trends within the city. Analyzing the demographic and housing factors that are associated with this change, we found that residents who are less affluent and who rent their homes were associated with earlier changes. This suggests that residents who are less knowledgeable about the infrastructure of their residence or who may feel more vulnerable, are quicker to be concerned with environmental dangers in their homes. Additionally, we find that households with residents over 60 are also associated with earlier change dates. This may indicate a particular advantage of having older members in a household who potentially have the insight and forethought to take action and test for potential environmental hazards.

While our conclusions do not constitute causal results, they demonstrate that our change surface model can provide unique insight to real world problems. By modeling change surfaces and characterizing the various functional regimes we hope this method can enable policy makers and scientists to design interventions that account for the heterogeneous complexity of human behaviors.

1.1 Literature review

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).

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 times $t \leq t_0$, 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.

1.2 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 allows us to analytically optimize the entire 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. Additionally, we apply the model to requests for lead testing kits in New York City over the past two years. Our results illustrate distinct spatial patterns in increased concern about lead-tainted water.

1.3 Outline

In the remainder of the paper, section 2 provides background on Gaussian processes. Section 3.1 describes our change surface model including the weighting, warping, and kernel functions. Section 3.2 introduces our novel algorithm for approximating the log determinant of additive kernels. Section 3.3 details our initialization procedure including our new approach for spectral mixture hyperparameter initialization. Section 4 describes our numerical and real-world experiments. Finally, we conclude with summary remarks in section 5.

2 Gaussian Processes

Given data $(\boldsymbol{y}, \boldsymbol{x})$, where $\boldsymbol{y} = \{y_1 ... y_n\}$ are outputs or response variables, and $\boldsymbol{x} = \{x_1 ... x_n\}, x_i \in R^D$, 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 $\boldsymbol{y} = f(\boldsymbol{x}) + \epsilon$, $f(x) \sim GP(m, k)$, $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon)$. A Gaussian process is a nonparametric prior over functions completely specified by mean and covariance functions:

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')) \tag{1}$$

$$m(x) = \mathbb{E}[f(x)] \tag{2}$$

$$k(x, x') = \text{cov}(f(x), f(x')) \tag{3}$$

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, $\boldsymbol{x}$.

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

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

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

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

3 Methodology

3.1 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, $w_i(x) : R^D \to R^1$, 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, $\sigma(z) : R^1 \to [0, 1]$, 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

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

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.

$\sigma(w_1(x)) \dots \sigma(w_r(x))$ induce non-stationarity 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, non-stationary 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 $\boldsymbol{x}$ and $\boldsymbol{y}$, 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 $\boldsymbol{x}$ and $\boldsymbol{y}$. 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.

3.1.1 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 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

which maps the D dimensional input x to an m dimensional feature space,

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

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

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

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 $\boldsymbol{x}$ we know that,

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

Equivalently, we can define parameters a such that,

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

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

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

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

allowing us to sample from $w(x)$ with a finite sum of RKS features. Initialization of hyperparameters σ_0 and Λ is discussed in Section 3.3.

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. Additionally, specifying the functional form of $w(x)$ does not require prior knowledge about if, where, or how rapidly changes occur.

3.1.2 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 $\tilde{x} \in R^P$ 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 3.3 which we hope can enable other models to exploit expressive kernels as well.

3.2 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 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 $\beta_1, \ldots, \beta_n$,

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

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 4.1). 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

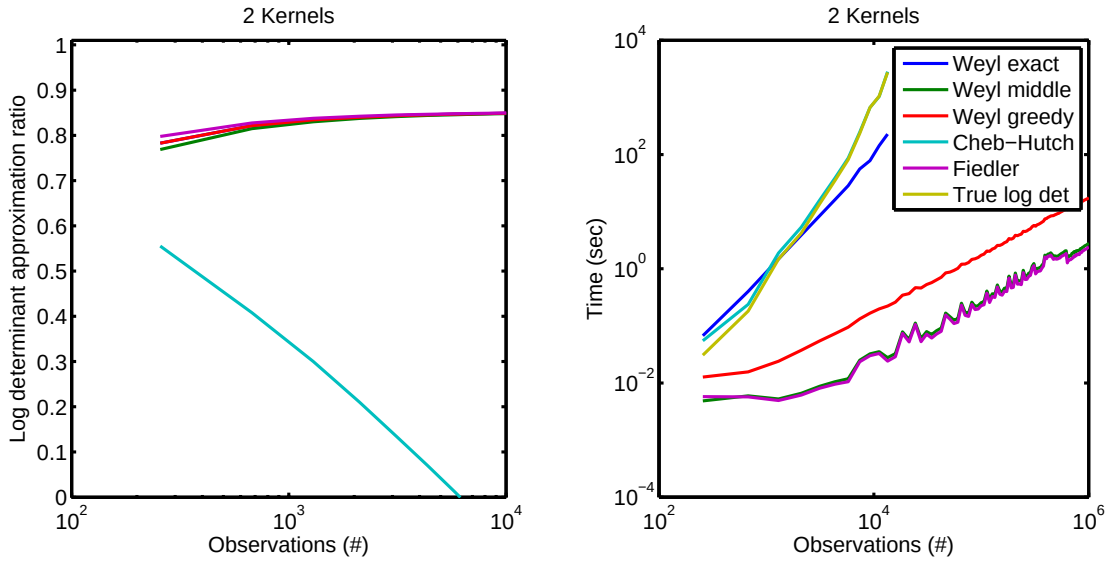

Figure 1: Left plot shows the ratio of approximations to the true log determinant of 2 additive kernels. Right plot shows the time to compute each approximation and the true log determinant of 2 additive kernels.

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. Finally, the Chebyshev-Hutchinson method scales poorly in our case due to expensive matrix-matrix multiplications required to construct a full K matrix.

3.3 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. Below we present a method where we first initialize the $w(x)$ RKS features and then use those values in a novel initialization method for the spectral mixture kernels.

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

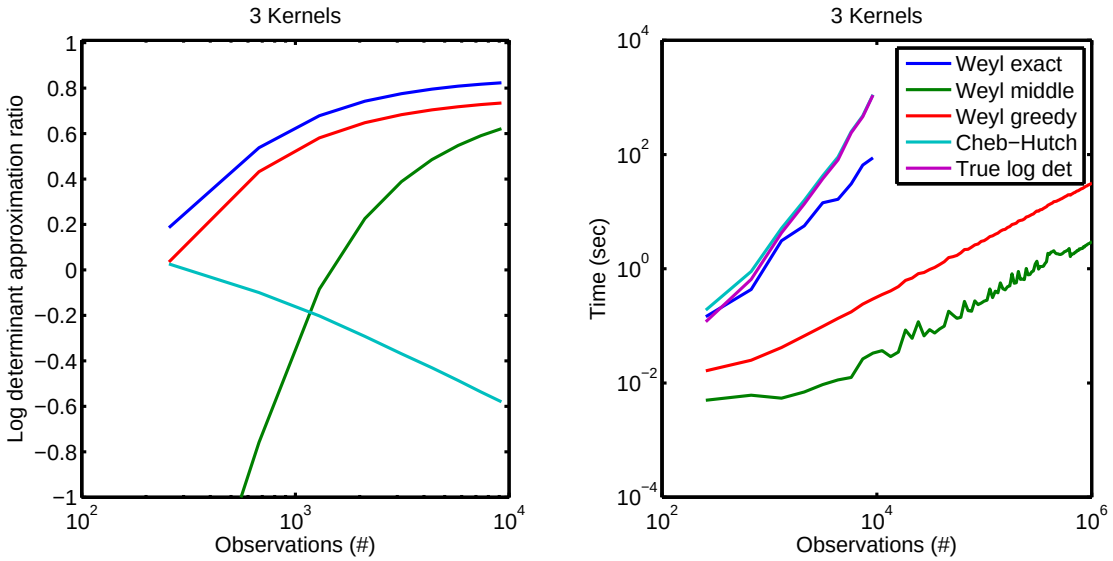

Figure 2: Left plot shows the ratio of approximations to the true log determinant of 3 additive kernels. Right plot shows the time to compute each approximation and the true log determinant of 3 additive kernels.

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 1 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 1 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 (see Section 3.1.1) g

number of times. We set reasonable values for hyperparameters in those 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 yields 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 yields the maximum marginal likelihood. Finally, we optimize all the resulting parameters until convergence.

In order to initialize the spectral mixture kernels, we use the initialized $w(x)$ from above to 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 2 Initialize spectral mixture kernels

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

4 Results

We demonstrate the model on multidimensional numerical simulations as well as real world data. There do not exist standard datasets for evaluating multidimensional or spatio-temporal changepoint models. For example, Majumdar et al. (2005) used simulations to demonstrate the effectiveness of their model. Therefore, we illustrate our method on a standard 1D changepoint dataset frequently used in the changepoint literature. Additionally we apply our model to two real world multidimensional datasets.

4.1 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. As emphasized in Section 3.3, successful convergence is dependent on reasonable initialization. Therefore, we use $g = 100$ and $h = 20$ for Algorithm 1. 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.

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 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 non-stationary co-

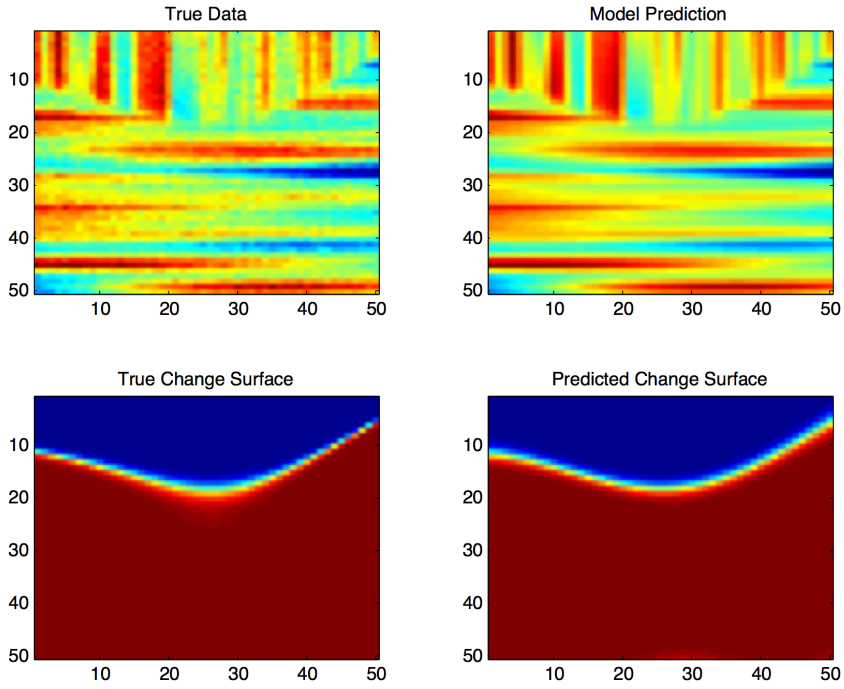

Figure 3: Numerical data experiment. The top-left depicts the data; the bottom-left shows the true change surface with the range from blue to red depicting $\sigma(w_1(x))$. The top-right depicts the predicted output; the bottom-right shows the predicted change surface.

variance 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.

4.2 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,

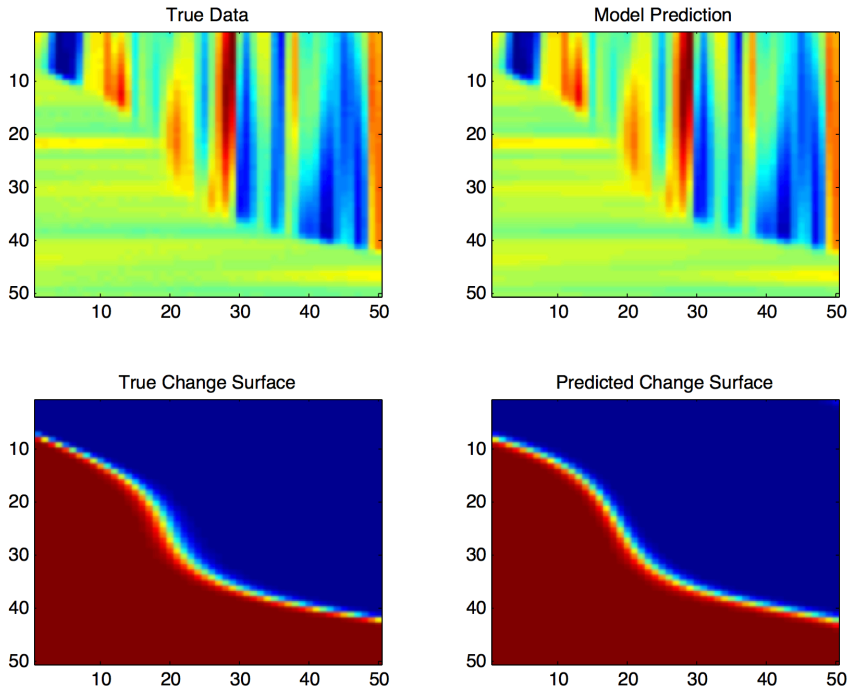

Figure 4: Numerical data experiment. The top-left depicts the data; the bottom-left shows the true change surface with the range from blue to red depicting $\sigma(w_1(x))$. The top-right depicts the predicted output; the bottom-right shows the predicted change surface.

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.

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

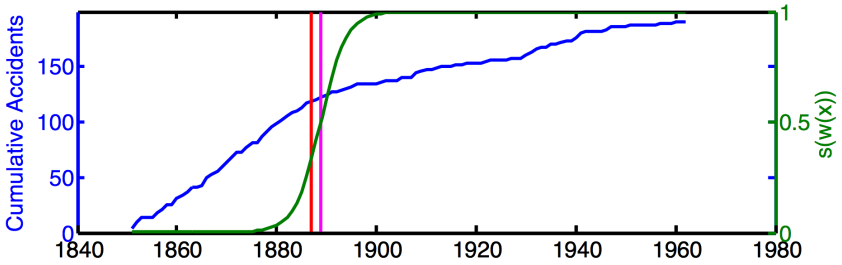

Figure 5: British coal mining accidents from 1851 to 1962. The blue line depicts cumulative annual accidents, the green line plots $\sigma(w(x))$, the vertical red line marks the Coal Mines Regulation Act of 1887, and the vertical magenta line indicates $\sigma(w_1(x)) = 0.5$.

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.

4.3 United States Measles Data

Measles was nearly eradicated in the United States following the introduction of the measles vaccine in 1963. However, due to the vast geographic, ethnic, bureaucratic, and socio-economic heterogeneity in the United States we may expect differential effectiveness of the vaccination program, particularly in its initial years. We analyze monthly incidence data for measles from 1935 to 2003 in each of the continental United States and the District of Columbia. Incidence rates per 100,000 population based on historical population estimates are 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

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

of each state and one temporal dimension.

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 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 United States. 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 illustrate the spatial variation in the change surface midpoint by shading states with an early midpoint in red and states with later midpoint in blue. We discover that there is an approximately 6 year difference in midpoint between states with California being the earliest and North Dakota being the latest.

In Figure 8 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. We illustrate the variation in slope by shading states with the flatter change regions in red and the steeper change regions in blue. Here we find that some states had approximately twice the rate of change as others, with Arizona being the flattest and Maine being the steepest.

These variations in the change surface indicate that the measles vaccine may have affected states heterogeneously over space and time. In order to better understand these dynamics we considered demographic information that may have contributed to differences in measles vaccine program implementation and effectiveness. Specifically we examined potential factors influencing the shift between the two regimes, $\sigma(w(x_{state})) = 0.5$, which we

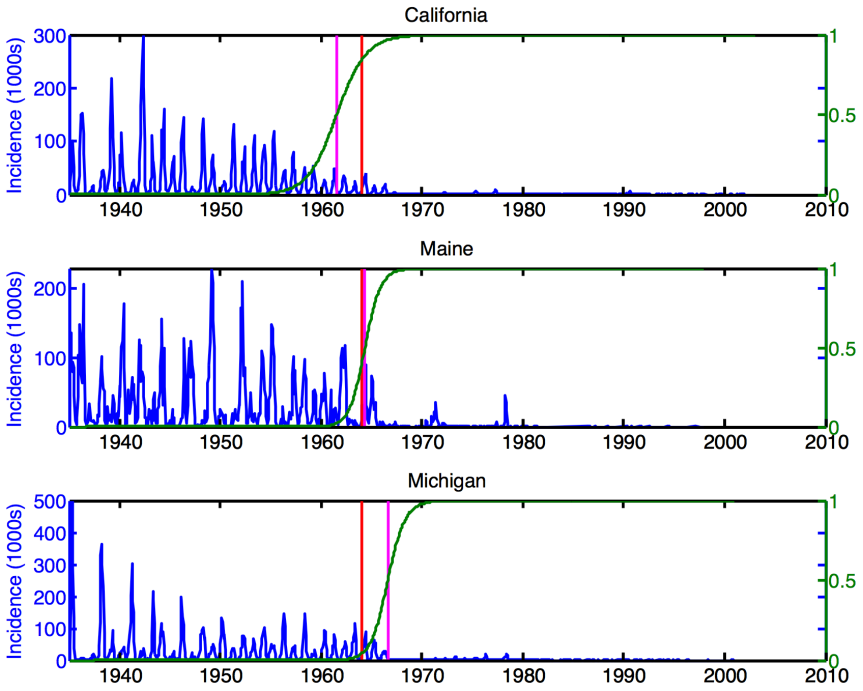

Figure 6: Measles incidence levels from 3 states, 1935 - 2003. The green line plots $\sigma(w(x_{state}))$, the vertical red line indicates the vaccine in 1963, and the magenta line indicates $\sigma(w(x_{state})) = 0.5$.

refer to henceforth as the change date. Since the change surface shifts primarily during the 1960's, we consider data only from that decade, averaging among years when data is available for multiple years. These factors included average annual birth rates, death rates of four age segments in the population, and absolute numbers of four age segments of the population in each state. Since measles is often contracted by children and people are rarely diagnosed for the disease twice in their life (it is a permanently immunizing disease), previous literature have shown that birth rates and the size of a young non-immune population is important for understanding the pre-vaccination dynamics of measles vaccines (Earn et al., 2000). We also consider average annual per capita income, median household income, and household income inequality for each state. Additionally, we use an average of the 1962 and 1967 censuses to compute the number of hospital and health workers in each state per population as a proxy for the size of the state government bureaucracy dedicated to implementing health policy. While this is an imperfect proxy as it neglects private hospitals and doctors, it does include health care workers focussed on immuniza-

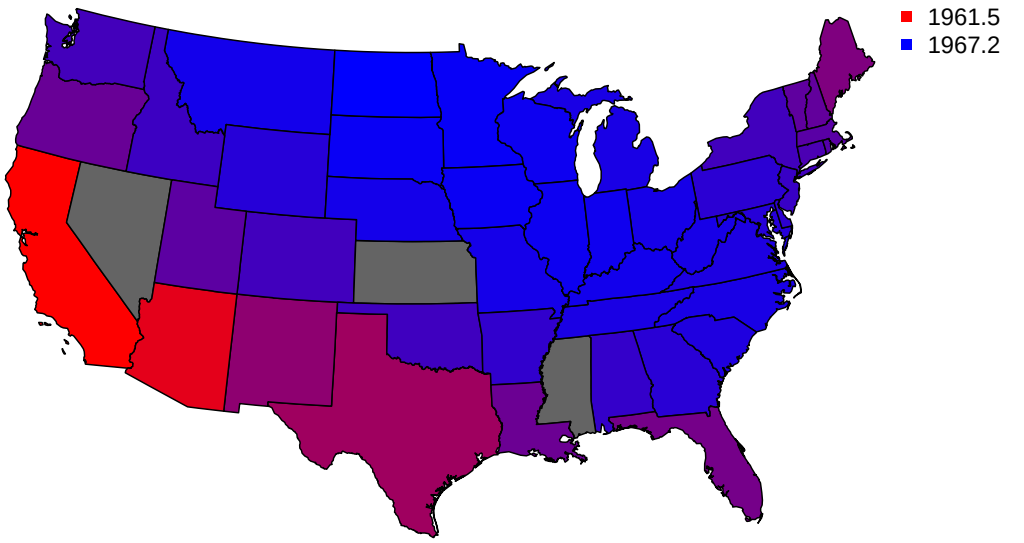

Figure 7: US states colored by the date where $\sigma(w(x_{state})) = 0.5$. Red indicates earlier dates, with California being the earliest. Blue indicates later dates, with North Dakota being the latest. Grayed out states were missing in the dataset.

tion. Unfortunately we were not able to separate out the number of employees concerned specifically with immunization. Finally, we also consider the slope of the change surface and average temperature in each state. Data were derived from historical census data (Census Bureau, 1999).

The results of a linear regression over all factors can be seen in Table 3. Two variables were statistically significant at a p-value < 0.05: the Gini coefficient of annual family income per state and the slope of the change surface from $\sigma(w(x_{state})) = 0.25$ to 0.75. Both of these features have relatively large, positive coefficients. This suggests that wider family income inequality is associated with later dates of switching to the post-vaccine regime. One potential explanation of this phenomenon may be that states with higher Gini coefficients may have had large socio-economically depressed communities as well as substantial rural populations. Inoculation and vaccination education may have been more difficult in those communities and regions, thus delaying the change date in those states. For example, Arkansas, Alabama, Kentucky, and Tennessee are all relatively rural states and have among the highest Gini coefficients. These states all have relatively late change dates sometime in 1966. Another interesting example is the District of Columbia, which

Table 3: Results from a linear regression to the United States measles change date, $\sigma(w(x_{state})) = 0.5$.

	Dependent variable:
	Change date
Average date rate 0-4	−228.924
	(687.066)
Average date rate 5-9	8,928.639
	(6,153.325)
Average birth rate	−0.210
	(0.188)
Gini of family income	32.317*
	(12.071)
Per capita income	0.00001
	(0.0002)
Slope of change surface	37.913**
	(8.976)
Gov't health and hospitals employees per population	326.952
	(165.077)
Population 0-4	−0.00002
	(0.00003)
Population 5-9	0.00002
	(0.00003)
Average temperature (°F)	0.025
	(0.041)
Constant	1,946.783**
	(7.614)
Observations	46
R^2	0.618
Adjusted R^2	0.446
Note:	*p<0.05; **p<0.01

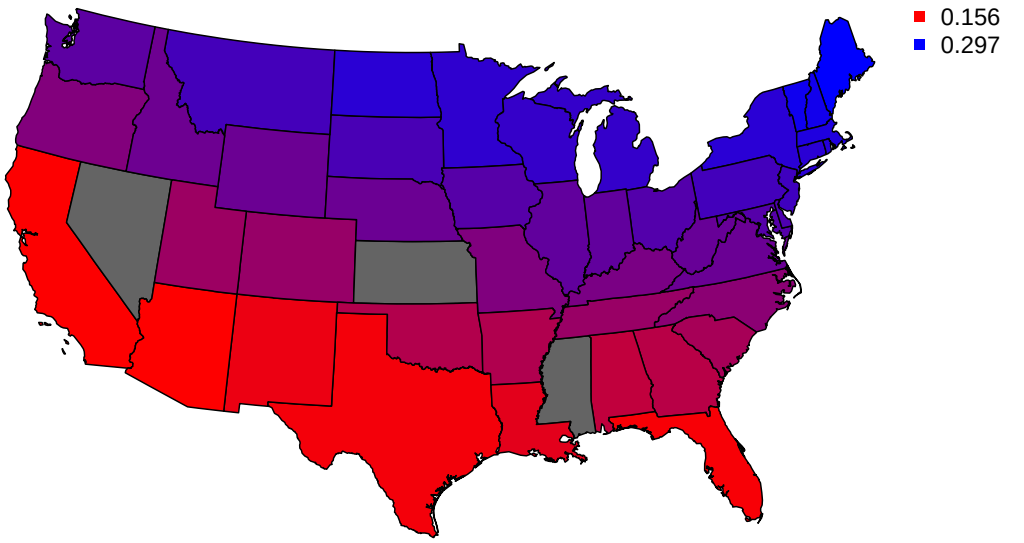

Figure 8: US states colored by the slope of $\sigma(w(x_{state}))$ from 0.25 to 0.75. Red indicates flatter slopes, with Arizona being the lowest. Blue indicates steeper slopes, with Maine being the highest. Grayed out states were missing in the dataset.

had the highest Gini coefficient. Although Washington DC is an urban center, it has long been an area of poverty and substandard local government, which may have contributed to its late change.

The correlation between change slope and change date suggests that states with later changes transition more quickly from the pre-vaccine regime to the post-vaccine regime. The steeper change slope may be due to other states already having inoculated their residents. Fewer measles cases nationwide could have enabled states with later change dates to more effectively contain the disease in their borders.

While this analysis does not provide conclusive results about underlying causal mechanisms, it suggests that further scientific research is warranted to understand the political and demographic factors that contributed to differential effectiveness in the early years of the measles vaccine program. Our conclusions indicate that future vaccination programs should particularly consider how to quickly and effectively provide vaccinations to rural areas and provide additional resources to socioeconomically disadvantaged communities.

4.4 New York City Lead Data

In recent years there has been heightened concern about lead-tainted water in major U.S. metropolitan areas. Specifically, since 2015, concerns about lead poisoning in Flint, Michigan's water supply have garnered national attention including Congressional hearings. Similar, if potentially less extensive, lead contamination issues have been reported in a spate of United States cities such as Cleveland, OH, New York City, NY, and Newark, NJ (Editorial Board, 2016). Lead concerns in New York City have focused on lead-tainted water in schools and public housing projects, prompting reporting in some local and national media (Gay, 2016).

In order to understand the evolving dynamics of New York City residents' concern about lead-tainted water, we analyzed requests for residential lead testing kits in New York City. These kits can be freely ordered by any resident of New York City and allow individuals to test their household's water for elevated levels of lead (New York, 2016). We considered weekly requests for each zip code in New York City from January 2014 through April 2016. This provides a proxy for measuring the concern about lead tainted water. Figure 9 shows the aggregated requests over the entire city for lead testing kits during the observation period. It could be argued that this is an imperfect reflection of citizen concern since is unlikely that a household will request more than one testing kit within a relatively short period of time. Thus a reduction in requests may be due to saturation in demand for kits rather than a decrease in concern. However, we contend that since there were only 28,057 requests for lead testing kits over the entire observation period, and New York City contains approximately 3,148,067 households, there is a substantial pool of households in New York City that are able to signal their concern through requesting a lead testing kit (Census Bureau, 2014a).

While there is a distinct uptick in requests for kits towards the middle and end of the observation period, unlike the coal mining and measles examples there is no known ground truth change point. We apply the change surface model with two latent functions, spectral mixture kernels, and $w(x)$ defined by 5 RKS features.

The model suggests that residents' concerns about lead tainted water had distinct spatial and temporal variation. In Figure 10 we depict the midpoint, $\sigma(w(x_{zip})) = 0.5$, for each zip code. We illustrate the spatial variation in the change surface midpoint by shading states with an early midpoint in red and states with later midpoint in blue. Regions in Staten Island and Brooklyn experienced the earliest midpoints, with Bulls Head in Staten

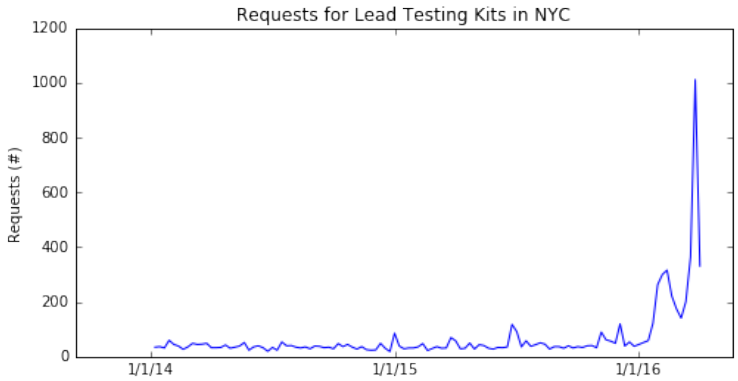

Figure 9: Requests for residential lead testing kits in New York City aggregated at a weekly level across the entire city.

Island (zip code 10314) being the first area to reach $\sigma(w(x_{zip})) = 0.5$ and New Hyde Park at the eastern edge of Queens (zip code 11040) being the last. The model detects certain zip codes changing in mid to late 2014, which somewhat predates the national publicity of the Flint water crisis. However, most zip codes have change dates sometime in 2015.

In Figure 11 we depict the change surface slope from $\sigma(w(x_{zip})) = 0.25$ to $\sigma(w(x_{zip})) = 0.75$ for each zip code to estimate the rate of change. We illustrate the variation in slope by shading states with the flatter change regions in red and the steeper change regions in blue. The flattest change surface occurred in Mariner's Harbor in Staten Island (zip code 10303) while the steepest change surface occurred in Woodlawn Heights in the Bronx (zip code 10470). We find that some zip codes had approximately four times the rate of change as others.

These variations in the change surface indicate that the concerns about lead-tainted water may have varied heterogeneously over space and time. In order to better understand these patterns we considered demographic and housing characteristics that may have contributed to differential concern among residents in New York City. Specifically we examined potential factors influencing the change date between the two regimes, where $\sigma(w(x_{zip})) = 0.5$. All data were taken from the 2014 American Community Survey 5 year average at the zip code level (Census Bureau, 2014b). Factors we considered included information about residents such as race of householder, education of householder, whether the householder was the home owner, previous year's annual income of household, number of people per household, and whether a minor or senior lived in the household. Additionally,

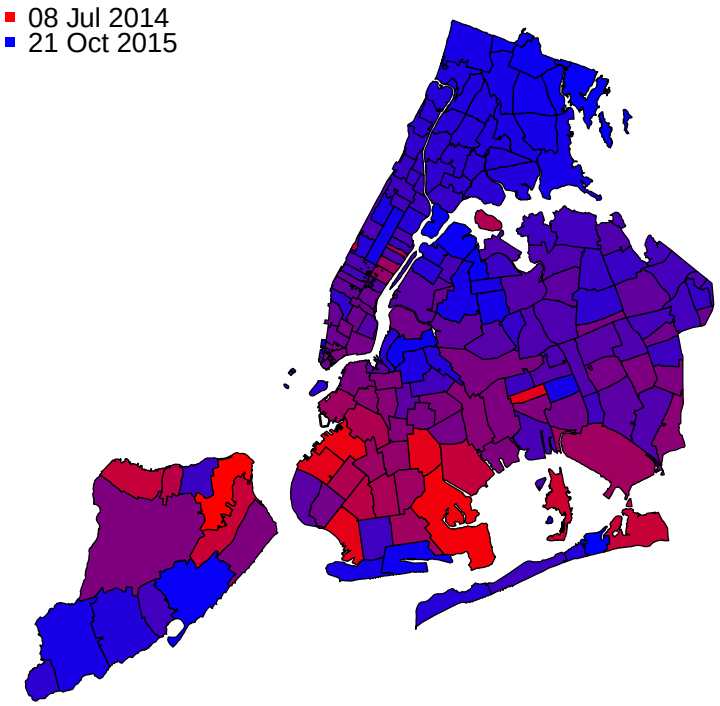

Figure 10: NYC zip codes colored by the date where $\sigma(w(x_{zip})) = 0.5$. Red indicates earlier dates, with Bulls Head in Staten Island being the earliest. Blue indicates later dates, with New Hyde Park at the eastern edge of Queens being the latest.

we considered information about when the homes were built. Finally, we also considered the slope of the change surface in that household's zip code.

Results of a linear regression over all factors can be seen in Table 4. Three variables were significantly positively correlated with change dates: median annual household income for the previous year, percentage of owner occupied households, and slope of the change surface. Since none of the education variables were significantly correlated we believe that the correlation with income is not a proxy for education. Instead, people with lower incomes may tend to live in housing that is less well maintained, or at least they fear is less well maintained. Similarly, renter households (the inverse of owner occupied households) may be less knowledgable about the infrastructure and plumbing in their homes. Thus renters and less affluent households may require less "activation energy" to request lead testing kits when faced with possible environmental hazards. The positive correlation between

Table 4: Results from a linear regression to the NYC lead change date, $\sigma(w(x_{state})) = 0.5$

	Dependent variable:
	Change date
Median annual household income	0.00005** (0.00002)
House built after 2010 (%)	0.259 (2.798)
House built 2000-09 (%)	0.320 (2.801)
House built 1980-099 (%)	0.327 (2.798)
House built 1960-79 (%)	0.375 (2.799)
House built 1940-59 (%)	0.377 (2.796)
House built before 1939 (%)	0.390 (2.798)
Householder African American (%)	−0.009 (0.081)
Householder Native American (%)	−0.490 (0.749)
Householder Asian (%)	0.057 (0.089)
Householder Pacific Islander (%)	−0.975 (1.993)
Householder other race (%)	−0.010 (0.048)
Householder Hispanic (%)	−0.009 (0.071)
Householder White (%)	−0.055 (0.083)
Education less than high school	−4.459 (3.438)
Education high school equivalent	−4.569 (3.441)
Education some college	−4.302 (3.437)
Education at least college	−4.519 (3.442)
Household owner occupied (%)	7.665*** (2.437)
Household average size	−4.267 (2.969)
Family average size	2.315 (2.271)
Household with member 18 or younger (%)	0.034 (0.073)
Household with member 60 or older (%)	−0.212*** (0.071)
Household with one member (%)	−0.111 (0.076)
Household with one member 65 or older (%)	0.152 (0.154)
Slope of change surface	7.655*** (0.149)
Constant	2,831.763*** (434.697)
Observations	174
R^2	0.979
Adjusted R^2	0.975
Residual Std. Error	2.381 (df = 147)
F Statistic	260.107*** (df = 26; 147)
Note:	*p<0.1; **p<0.05; ***p<0.01

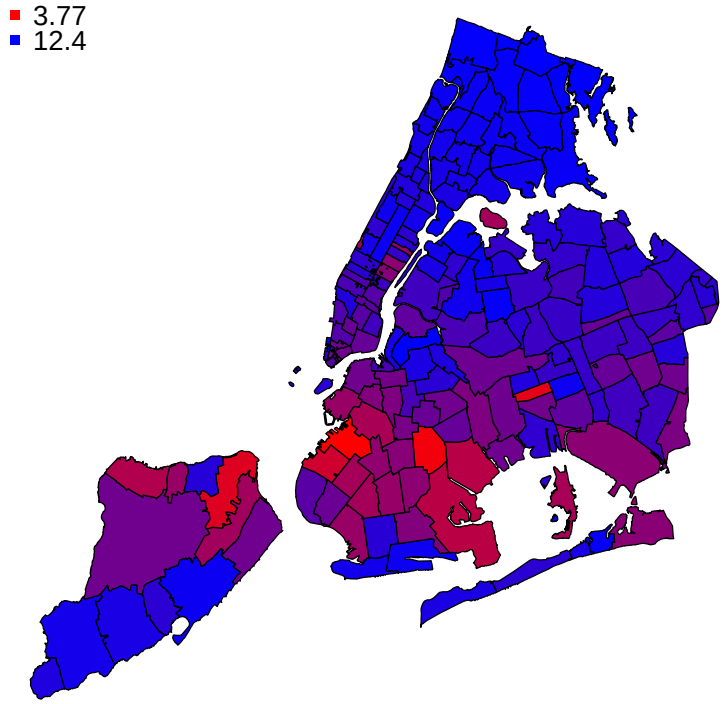

Figure 11: NYC zip codes colored by the slope of $\sigma(w(x_{zip}))$ from 0.25 to 0.75. Red indicates flatter slopes, with Mariner's Harbor in Staten Island being the flattest. Blue indicates steeper slopes, with Woodlawn Heights in the Bronx being the steepest.

change date and change slope is evident from a visual inspection of Figures 10 and 11. This relation indicates that in zip codes that changed later, their changes were relatively quicker.

The percentage of households with persons 60 and older was significantly negatively correlated with change date. While the coefficient of this variable is proportionally smaller than the previous variables, it suggest that households with older people may be more sensitive to environmental dangers. Indeed, it may be that these older citizens have the time to spend reading the latest news reports and concerning themselves with issues such as lead-tainted water, or else they have the historical memory to recall the dangers such issues have posed in the past.

This analysis indicates that more educational outreach from utility providers and the New York City Department of Environmental Protection could help address residents con-

cerns about lead-tainted water. Additionally, it suggests an information disparity between renters and owner-occupiers that may be of interest to policy makers. Finally, it shows that in certain cases there may be a particular advantage of older members in a household who potentially have the insight and forethought to take action and test for potential environmental hazards.

Beyond the statistical analysis of demographic data, we also qualitatively examined media coverage related to the Flint water crisis as detailed by the Flint Water Study (Water Study, 2015). While a few articles and news reports were reported in 2014, the vast majority began in 2015. The increased rate and national scope of this coverage in 2015 and 2016 may explain why zip codes with later change dates shifted more rapidly. Additionally, it may be that residents with lower incomes identified earlier with those in Flint and thus were more concerned about potentially contaminated water than their more affluent neighbors.

5 Conclusion

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. The analysis of measles data in the United States and requests for lead testing kits in New York City demonstrate how the model can be used to yield scientifically and policy 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 current method can be extended to automatically determine the number of latent functions using a automatic modeling discovery approach such as Lloyd et al. (2014).

Acknowledgements

Thank you for the invaluable advice and assistance of Andrew Wilson, Hannes Nickisch, and Seth Flaxman. Also, thank you to Proma Paul and Linda Leqi Huang for compiling the population statistics for the measles regression analysis.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE 1252522 and the National Science Foundation award No. IIS-0953330.

References

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

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

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

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

Census Bureau, U. S. (1999), "United States Historical Census Data," https://www.census.gov/hhes/www/income/data/historical/state/, accessed: 2016-4-10.

— (2014a), "American Community Survey 1-Year Estimates," http://factfinder.census.gov/, accessed: 2016-4-10.

— (2014b), "American Community Survey 5-Year Estimates," http://factfinder.census.gov/, accessed: 2016-4-10.

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*, 999–1018.

Earn, D. J. D., Rohani, P., Bolker, B. M., and Grenfell, B. T. (2000), "A Simple Model for Complex Dynamical Transitions in Epidemics," *Science*, 287, 667–670.

Editorial Board, T. (2016), "Poisoned Water in Newark Schools," *New York Times*.

Fiedler, M. (1971), "Bounds for the determinant of the sum of hermitian matrices," *Proceedings of the American Mathematical Society*, 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*, ACM, pp. 345–352.

Gay, M. (2016), "Elevated Levels of Lead Found in Water of Some Vacant Public-Housing Apartments," *Wall Street Journal*.

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)*, pp. 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, 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, 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, 55–81.

New York, C. o. (2016), "Water Lead Test Kit Request," http://www1.nyc.gov/nyc-resources/service/1266/water-lead-test-kit-request, accessed: 2016-4-10.

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*, 85–89.

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

Rasmussen, C. and Williams, C. (2006), *Gaussian Processes for Machine Learning*, MIT Press.

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

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

Saatçi, Y. (2012), "Scalable inference for structured Gaussian process models," Ph.D. 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)*, pp. 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, 2152.

Water Study, F. (2015), "Flint Water Study: Articles in the Press," http://flintwaterstudy.org/articles-in-the-press/, accessed: 2016-4-10.

Weyl, H. (1912), "Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung)," *Mathematische Annalen*, 71, 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*, pp. 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)*, pp. 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*, pp. 3626–3634.

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