Change modeling for understanding our world and the counterfactual one(s)

William Herlands

Committee: Daniel Neill (NYU)
Andrew Gordon Wilson (NYU)
Akshaya Jha
Roni Rosenfeld
Leman Akoglu

Machine Learning Department & H. John Heinz III College
Carnegie Mellon University

This dissertation is submitted for the degree of
Doctor of Philosophy in Machine Learning and Public Policy

Pittsburgh, PA April 2020

"A thinker sees his own actions as experiments and questions – as attempts to find out something. Success and failure are for him answers above all."
- Friedrich Nietzsche

Dedicated to my family, friends, and mentors who inspire curiosity and creativity, particularly Natalia and Lincoln.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original. This dissertation is my own work achieved in collaboration with far more intelligent collaborators.

William Herlands
April 2020

Acknowledgements

This thesis would have been impossible without my advisors, Daniel Neill and Andrew Gordon Wilson, who provided invaluable guidance over the past six years. I would also like to acknowledge the tremendous assistance of my collaborators, Maria De-Arteaga, Artur Dubrawski, Seth Flaxman, Edward McFowland III, Hannes Nickisch, and Wilbert van Panhuis. Finally I would like to thank my additional thesis committee members Akshaya Jha, Roni Rosenfeld, and Leman Akoglu, as well as David Choi for chairing all my official presentations, for their insights, time, and patience.

During my PhD I was the beneficiary of multiple grants including the NSF GRFP and an ARCS Foundation Fellowship. I am deeply indebted to those organizations for their tremendous dedication to furthering cutting edge research.

Abstract

Detecting, analyzing, and modeling changes provide essential information for understanding scientific processes and human behavior. While change analysis is fundamental in machine learning and statistics, many standard models are limited in expressiveness or make unrealistic simplifying assumptions. This thesis focuses on two interrelated elements of change analysis.

First, we provide rich characterization of changes by developing new methods for modeling complex changes and for detecting anomalous patterns in real world data. In order to characterize a change we automatically model the "null" regions of stability in the data and identify where "alternative" regions of change or anomalies exist. By modeling how the alternative regions differ or evolve from the null regions, we show that we are able to use that information for scientific discovery and for early event detection.

Second, we consider causal and counterfactual inference by exploiting changes to uncover the generative structure of data. By isolating changes in data we can reason about what would have occurred in the absence of a change. Such reasoning enables us to predict the counterfactual world and estimate the causal impact of certain variables or interventions in a data set.

Using a dozen different public interest data sets we employ our methods to characterize changes and identify causal mechanisms that can provide scientific and policy relevant insights. Specifically, we concentrate on health policy and urban data, much of which exhibit distinct spatial and demographic patterns. The data we explore includes measles incidence, health insurance usage rates, water lead testing, requests for municipal services, urban opioid deaths, weather related damage in urban neighborhoods, urban school absenteeism, and police traffic stops.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

A season is set for everything, a time for every experience under heaven:
A time for being born and a time for dying, A time for planting and a time for uprooting;
A time for slaying and a time for healing, A time for tearing down and a time for building up;
A time for weeping and a time for laughing, A time for wailing and a time for dancing;
A time for throwing stones and a time for gathering stones, A time for embracing and a time for shunning embraces;
A time for seeking and a time for losing, A time for keeping and a time for discarding;
A time for ripping and a time for sewing, A time for silence and a time for speaking;
A time for loving and a time for hating; A time for war and a time for peace.

Ecclesiastes 3:1-8

Everything changes. Yet Ecclesiastes emphasizes that as humans we often believe that our world is somehow fixed. We're surprised not only by political revolution but also by highly predictable changes from Summer to Fall to Winter. And while statistical methods exist to detect and use changes in data distributions, these methods tend to model relatively simple changes or else require substantial expert human oversight. The potential for modeling complex changes or automatically using changes in data to deeply understand a system are vastly under-explored.

This thesis pays homage to changes. The fundamental contention throughout these chapters is that detecting, analyzing, and modeling changes provide essential information for understanding scientific processes and human behavior.

Change themes: Within the overarching theme of change detection we concentrate on two major elements:

- Characterizing changes by developing new methods for modeling complex changes and for detecting anomalous patterns in complex data. In order to characterize a change we automatically model the "null" regions of stability in the data and identify where "alternative" regions of change or anomalies exist. By modeling how the alternative regions differ or evolve from the null regions, we show that we are able to use that information for scientific discovery or for early event detection.

- Causal and counterfactual inference by exploiting changes to uncover the generative structure of data. By isolating changes in data we can reason about what would have occurred in the absence of a change. Such reasoning enables us to predict the counterfactual world and estimate the causal impact of certain variables or interventions in a data set.

Methodological themes: Throughout this thesis we will use and innovate on two statistical machine learning methodologies:

- Gaussian processes. GPs are particularly suited for change modeling since they allow for quite general modeling of data with closed form expressions for inference and likelihood estimation. GPs are a Bayesian non-parametric technique that provide a natural means for encoding expert insight, or previously known characteristics of the data. With respect to application domains, discussed in the paragraph below, GPs are a useful model for spatiotemporal data since the covariance function can automatically learn correlations across multiple dimensions. In public policy data we are also often faced with missing or partially incomplete datasets, which are naturally handled by GPs without any need for special modification. We provide a more detailed introduction to GPs in Section 1.2.

- Subset scanning. Identifying where changes or anomalies exist is a challenging task because we must both model the null data where no changes exist while also searching for the alternative regions which exhibit a significant change in the data distribution. By searching over constrained subsets of data, subset scanning provides an efficient and formalized mechanism to discover changes in multidimensional data and can be adapted to datasets with either real-valued or categorical variables and outputs.

Application themes: While this thesis focuses on developing novel statistical machine learning methods to identify and exploit changes, we have a specific interest in application areas of public importance. Change analysis is particularly important for understanding scientific processes and public policy. In these fields it is often practically or morally difficult

to run highly controlled experiments so researchers are left to use observational data to understand a system. By focusing on regions of changes we can gain insight into the often hidden mechanisms undergirding such systems. Specifically we concentrate on health care data (such as historical epidemiological trends in Section 2.4.4) and urban policy data (such as New York City 311 data in Section 4.4). Much of the data is spatiotemporal, whose substantial correlations across dimensions adds complexity to our endeavor. While readers with an exclusive interest in statistical methods may be tempted to ignore these application sections, we encourage them to not be so hasty. The application of machine learning methods to truly real data (rather than toy "real world" data) is an essential aspect of verifying the utility of any new method. We spend considerable time in this thesis analyzing experimental results to demonstration of how our methods can provide real insight to stakeholders.

1.1 Outline

Each chapter in this thesis addresses some combination of the themes described above. Their purposes, methodologies, and applications build on one another and provide a general perspective for considering how to approach a variety of changes in different types of data.

Chapter 2: The thesis begins by challenging changepoints, a ubiquitous framework for modeling changes in data. Standard changepoint models are limited in expressiveness, often addressing unidimensional problems and assuming instantaneous changes. We introduce *change surfaces* as a multidimensional and highly expressive generalization of changepoints [63, 64]. We provide a model-agnostic formalization of change surfaces, illustrating how they can provide variable, heterogeneous, and non-monotonic rates of change across multiple dimensions. Additionally, we instantiate change surfaces by developing Gaussian Process Change Surfaces (GPCS) and we demonstrate the ability for massive scalability by introducing novel methods for additive non-separable kernels.

This chapter focuses on the themes of change characterization, GP methodology, and applies the methods to three public policy and public health datasets.

Chapter 3: Continuing with change surfaces, this chapter develops a framework for using change surfaces for counterfactual prediction. We discuss the assumptions necessary for the validity of these counterfactuals, including how they relate to the potential outcomes framework. Additionally, using Gaussian Process Change Surfaces we demonstrate counterfactual prediction with Bayesian posterior mean and credible sets.

This chapter focuses on the theme of casual inference, GP methodology, and applies the methods to a large public health dataset.

Chapter 4: In an effort to detect localized changes in spatiotemporal data, this chapter develops an anomalous pattern detection technique for highly correlated data [61]. Anomaly detection techniques often identify points on the peripheries of the data distribution, which is useful for tasks such as data cleaning or online monitoring for extreme points. Yet methods that separately consider the anomalousness of each individual data point have low detection power for subtle, emerging irregularities. Additionally, recent detection techniques based on subset scanning make strong independence assumptions and suffer degraded performance in correlated data. We introduce methods for identifying anomalous patterns in non-iid data by combining Gaussian processes with novel log-likelihood ratio statistic and subset scanning techniques. Our approaches are powerful, interpretable, and can integrate information across multiple data streams.

This chapter focuses on the themes of change characterization, both GP and subset scanning methodology, and applies the methods to four urban spatiotemporal data sets.

Chapter 5: Using anomalous pattern detection we develop the first statistical machine learning approach for automatically discovering regression discontinuity designs [62]. RDDs are a natural experiment setup often used in econometrics to infer causal treatment effects from observational datasets. Our method identifies interpretable, localized RDDs in arbitrary dimensional data and can seamlessly compute treatment effects without expert supervision.

This chapter focuses on the theme of casual inference, subset scanning methodology, and applies the methods to three data sets in education and health care.

Chapter 6: Continuing to focus at intersection of machine learning and econometrics we develop a framework for automatically discovering difference-in-differences (DD) in time series data. Our method extends the RDD search techniques from Chapter 5 to discover RDDs in heterogeneous categorical subsets of data. Specifically for DDs we apply this approach to temporal data and discover RDDs in time. Additionally we develop a novel approach for identifying control subsets for DDs in order to compute treatment effects.

This chapter focuses on the theme of casual inference, subset scanning methodology, and applies the methods to two public policy datasets.

Chapter 7: We conclude the thesis in this chapter with some observations about overarching methodological and conceptual themes that span multiple chapters.

1.2 Background on Gaussian processes

A number of chapters in this thesis rely heavily on Gaussian processes. As such we provide a brief review of Gaussian processes here which has relevance to work throughout the thesis. More details can be found in comprehensive works such as Rasmussen and Williams [113], Schölkopf and Smola [130], and MacKay [93].

Consider data, (x,y), where $x = \{x_1,\ldots,x_n\}, x_i \in \mathbb{R}^D$, are inputs or covariates, and $y = \{y_1,\ldots,y_n\}$, $y_i \in \mathbb{R}$ are outputs or response variables indexed by x. We assume that y is generated from x by a latent function with a Gaussian process prior (GP) and Gaussian noise. In particular,

$$y = f(x) + \varepsilon \tag{1.1}$$
$$f(x) \sim \mathscr{GP}(\mu(x), k(x,x')) \tag{1.2}$$
$$\varepsilon \sim \mathscr{N}(0, \sigma_\varepsilon^2) \tag{1.3}$$

A Gaussian process is a nonparametric prior over functions completely specified by mean and covariance functions. The mean function, $\mu(x)$, is the prior expectation of $f(x)$, while the covariance function, $k(x,x')$, is a positive semidefinite kernel that defines the covariance between function values $f(x)$ and $f(x')$.

$$\mu(x) = \mathbb{E}[f(x)] \tag{1.4}$$
$$k(x,x') = \mathrm{cov}(f(x), f(x')) \tag{1.5}$$

Any finite collection of function values is normally distributed $[f(x_1)...f(x_p)] \sim \mathscr{N}(\mu(x), K)$ where $p \times p$ matrix $K_{i,j} = k(x_i,x_j)$. Thus we can draw samples from a Gaussian process at a finite set of points by sampling from a multivariate Gaussian distribution. In this thesis we generally consider $\mu(x) = 0$ and concentrate on the covariance function. The choice of kernel is particularly important in Gaussian process applications since the kernel defines the types of correlations encoded in the Gaussian process. For example, a common kernel choice is a Radial Basis Function (RBF), also known as a Gaussian kernel,

$$k(x,x') = s^2 \exp[-(x-x')^T V^{-1}(x-x')/2] \tag{1.6}$$

where s^2 is the signal variance and V is a diagonal matrix of bandwidths. The RBF kernel implies that nearby values are more highly correlated. While this may be true in many applications, it would be inappropriate for data with significant periodicity. In such cases a periodic kernel would be more fitting. We consider more expressive kernel representations in

Section 2.3.1. This formulation of Gaussian processes naturally accommodates inputs x of arbitrary dimensionality.

Prediction with Gaussian processes Given a set of kernel hyperparameters, θ, and data, (x,y), we can derive a closed form expression for the predictive distribution of $f(x^*)$ evaluated at points x^*,

$$f(x^*)|\theta,x,y,x^* \sim \mathcal{N}\Big(k(x^*,x)[k(x,x)+\sigma_\varepsilon^2 I]^{-1}(y-\mu(x))+\mu(x^*), \\ k(x^*,x^*)-k(x^*,x)[k(x,x)+\sigma_\varepsilon^2 I]^{-1}k(x,x^*)\Big) \tag{1.7}$$

The predictive distribution provides posterior mean and variance estimates that can be used to define Bayesian credible sets. Thus Gaussian process prediction is useful both for estimating the value of a function at new points, x^*, and for deriving a function's distribution in the domain, x, for which we have data.

Learning Gaussian process hyperparameters In order to learn kernel hyperparameters we often desire to optimize the marginal likelihood of the data conditioned on the kernel hyperparameters, θ, and inputs, x.

$$p(y|\theta,x) = \int p(y|f,x)p(f|\theta)df \tag{1.8}$$

Thus we choose the kernel which maximizes the likelihood that the observed data is generated by the Gaussian process prior with hyperparameters θ. In the case of a Gaussian observation model we can express the log marginal likelihood as,

$$\log p(y|\theta,x) = -\frac{1}{2}\log|K+\sigma_\varepsilon^2 I| - \frac{1}{2}(y-\mu(x))^T(K+\sigma_\varepsilon^2 I)^{-1}(y-\mu(x)) + \text{constant} \tag{1.9}$$

Drawbacks However, solving linear systems and log determinants involving the $n \times n$ covariance matrix K incurs $\mathcal{O}(n^3)$ computations and $\mathcal{O}(n^2)$ memory, for n training points, using standard approaches based on the Cholesky decomposition [113]. These computational requires are prohibitive for many applications, particularly in scientific analysis and public policy — the foci of this thesis — where it is normal to have more than few thousand training points. Additionally, recent advances in GP scalability have often been developed at the expense of the expressiveness of the GP model. For example, Kronecker-based methods require data be distributed on a multi-dimensional lattice and are restricted to separable kernel functions across data dimensions.

As part of our work on change modeling, we address some challenges of working with GPs in modern datasets. In Chapter 2 we develop alternative inference procedures for additive kernels which enable scalable non-separable inference for GPs. Additionally, we develop novel initialization procedures for spectral mixture kernels which enable learning of more complex covariance functions.

Chapter 2

Change Surfaces for Generalized Change Modeling

2.1 Introduction

Detecting and modeling changes in data is critical in statistical theory, scientific discovery, and public policy. For example, in epidemiology, detecting changes in disease dynamics can provide information about when and where a vaccination program becomes effective. In dangerous professions such as coal mining, changes in accident occurrence patterns can indicate which regulations impact worker safety. In city governance, policy makers may be interested in how requests for health services change across space and over time[1].

Changepoint models have a long history in statistics, beginning in the mid-twentieth century, when methods were first developed to identify changes in a data generating process [108, 68]. The primary goal of these models is to determine if a change in the distribution of the data has occurred, and then to locate one or more points in the domain where such changes occur. While identifying these changepoints is an important result in itself, changepoint methods are also frequently applied to other problems such as outlier detection or failure analysis [118, 141, 80]. Different changepoint methods are distinguished by the diversity of changepoints they are able to detect and the complexity of the underlying data. The simplest models consider mean shifts between functional regimes [35, 83], while others consider changes in the covariance structure or higher order moments [82, 120, 76]. A *regime* is a particular data generating process or underlying function that is separated from other underlying processes or functions by changepoints. Additionally, there is a fundamental distinction between changepoint models that identify changes sequentially using online

[1] Published as Herlands et al. [64]

algorithms, and those that analyze data retrospectively to find one or more changes in past data [24, 33]. Finally, changepoint methods may be fully parametric, semi-parametric, or nonparametric [120, 54]. For additional discussion of changepoints beyond the scope of this chapter, readers may consider the literature reviews in Aue and Horváth [18], Ivanoff and Merzbach [74], and Aminikhanghahi and Cook [8].

Yet nearly all changepoint methods described in the statistics and machine learning literature consider system perturbations as discrete changepoints. This literature seeks to identify instantaneous differences in parameter distributions. The advantage of such models is that they provide definitive assessments of the location of one or more changepoints. This approach is reasonable, for instance, when considering catastrophic events in a mechanical system, such as the effect of a car crash on various embedded sensor readings. Yet the challenge with these models is that real world systems rarely exhibit a clear binary transition between regimes. Indeed, in many applications, such as in biological science, instantaneous changes may be physically impossible. While a handful of approaches consider non-discrete changepoints [e.g., 152, 155, 92] they still require linear, monotonic, one-dimensional, and, in practice, relatively quick changes. Existing models do not provide the expressiveness necessary to model complex changes.

Additionally, 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. [95] 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 [106] use a Bayesian onset-field process on a lattice to model the spatio-temporal distribution of human settlement on the Fiji islands. However, the models in these papers are limited to considering discrete changepoints.

2.1.1 Main contributions

In this chapter, we introduce *change surfaces* as expressive, multidimensional generalizations of changepoints. We present a model-agnostic formulation of change surfaces and instantiate this framework with scalable Gaussian process models. The resulting model is capable of automatically learning expressive covariance functions and a sophisticated continuous change surface. Additionally, we derive massively scalable inference procedures. Finally, we apply the proposed methods to a wide variety of numerical data and complex human systems. In particular, we:

1. Introduce change surfaces as multidimensional and highly flexible generalizations of changepoint modeling.

2. Introduce a procedure which allows one to specify background functions and change functions, for more powerful inductive biases and added interpretability.

3. Present the Gaussian Process Change Surface model (GPCS) which models change surfaces with highly flexible Random Kitchen Sink [111] features.

4. Develop massively scalable additive, non-stationary, non-separable kernels by using the Weyl inequality [150] and novel Kronecker methods. In addition we integrate our approach into the recent KISS-GP framework [154]. The resulting approach is the first scalable Gaussian process multidimensional changepoint model.

5. Describe a novel initialization method for spectral mixture kernels [151] by fitting a Gaussian mixture model to the Fourier transform of the data. This method provides good starting values for hyperparameters of expressive stationary kernels, allowing for successful optimization over a multimodal parameter space.

6. Demonstrate that the GPCS approach is robust to misspecification, and automatically discourages extraneous model complexity, leading to the discovery of interpretable generative hypotheses for the data.

7. Use GPCS for discovering and characterizing continuous changes in large observational data. We demonstrate our approach on a recently released public health dataset providing new insight that suggests how the effect of the 1963 measles vaccine may have varied over space and time in the United States. Additionally, we apply the model to requests for lead testing kits in New York City from 2014-2016. The results illustrate distinct spatial patterns in increased concern about lead-tainted water.

2.1.2 Outline

The chapter is divided into three main units.

Section 2.2 formally introduces the notion of change surfaces as a multidimensional, expressive generalization of changepoints and we discuss a variant of change surfaces in Section 2.2.1. The discussion of change surfaces in this unit is method-agnostic, and should be relevant to experts from a wide variety of statistical and machine learning disciplines. We emphasize the novel contribution of this framework to the general field of change detection.

Section 2.3 presents the Gaussian Process Change Surface (GPCS) as a scalable method for change surface modeling. We specify the GPCS model in Section 2.3.1. Scalable inference using novel Kronecker methods are presented in Section 2.3.2, and we describe a novel initialization technique for expressive Gaussian process kernels in Section 2.3.3.

Section 2.4 demonstrates GPCS on *out-of-class* numerical data and complex spatio-temporal data. We describe our numerical setup in Section 2.4.1 presenting results for posterior prediction and change surface identification. We present a one-dimensional application of GPCS on coal mining data in Section 2.4.2 including a comparison to state-of-the-art changepoint methods. Moving to spatio-temporal data, we apply GPCS to model requests for lead testing kits in New York City in Section 2.4.3 and discuss the policy relevant conclusions. Additionally, we use GPCS to model measles incidence in the United States in Section 2.4.4 and discuss scientifically relevant insights.

2.2 Change surfaces

In human systems and scientific phenomena 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 ingrained habits or trickle through a complex bureaucracy. The dynamics of these changes are non-trivial, with sophisticated distributions, rates, and intensity functions. Using expressive models to fully characterize such changes is essential for accurate predictions and scientifically meaningful results. For example, in the spatio-temporal domain, changes are often heterogeneously distributed across space and time. Capturing the complexity of these changes provides useful insights for future policy makers enabling them to better target or structure policy interventions.

In order to provide the expressive capability for such models, we introduce the notion of a *change surface* as a generalization of changepoints. We assume data are (x,y), where $x = \{x_1,\ldots,x_n\}, x_i \in \mathbb{R}^D$, are inputs or covariates, and $y = \{y_1,\ldots,y_n\}$, $y_i \in \mathbb{R}$, are outputs or response variables indexed by x. A change surface defines transitions between latent functions $f_1,\ldots,f_r$ defining r regimes in the data. Unlike with changepoints, we do not require that the transitions be discrete. Instead we define r warping functions $s(x) = [s_1(x),\ldots,s_r(x)]$ where $s_i(x) : \mathbb{R}^D \to [0,1]$, which have support over the entire domain of x. Importantly, these warping functions have an inductive bias towards $\{0,1\}$ creating a soft mutual exclusivity between the functions. We define the canonical form of a change surface as

$$\begin{aligned} y(x) &= s_1(x)f_1(x) + \cdots + s_r(x)f_r(x) + \varepsilon \\ &s.t. \\ &\sum_{i=1}^{r} s_i(x) = 1 \\ &s_i(x) \geq 0 \end{aligned} \tag{2.1}$$

where $\varepsilon(x)$ is noise. Each $s_i(x)$ defines how the coverage of $f_i(x)$ varies over the input domain. Where $s_i(x) \approx 1$, $f_i(x)$ dominates and primarily describes the relationship between x and y. In cases where there is no i such that $s_i(x) \approx 1$, a number of functions are dominant in defining the relationship between x and y. Since $s(x)$ has a strong inductive bias towards 1 or 0, the regions with multiple dominant functions are transitory and often the areas of interest. Therefore, we can interpret how the change surface develops and where different regimes dominate by evaluating each $s(x)$ over the input domain.

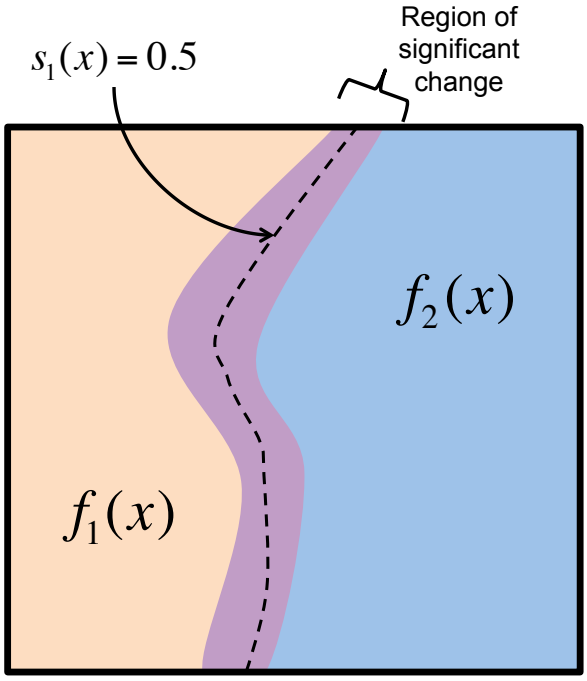


Fig. 2.1 Two-dimensional depiction of the change surface model where $f_1(x)$ is drawn in orange and $f_2(x)$ is drawn in blue. The region in purple depicts an area of transition between the two functions. The dashed line represents the domain where $s_1(x) = 0.5$.

Figure 2.1 depicts a two-dimensional change surface model where latent $f_1(x)$ is drawn in orange and latent $f_2(x)$ is drawn in blue. In those areas the first warping function, $s_1(x)$, is nearly 1 and 0 respectively. The region in purple depicts an area of transition between the two functions. We would expect that $s_1(x) \approx 0.5$ in this region since both latent functions are active.

In many applications we can imagine that a latent background function, $f_0(x)$, exists that is common to all data regimes. One could reparametrize the model in Eq. (2.1) by letting each latent regime be a sum of two functions: $f_0(x) + f_i(x)$. Thus each regime compartmentalizes into $f_0(x)$, a common background function, and $f_i(x)$, a regime-specific latent function. This

provides a generalized change surface model,

$$y(x) = f_0(x) + s_1(x)f_1(x) + \cdots + s_r(x)f_r(x) + \varepsilon(x). \tag{2.2}$$

Change surfaces can be considered particular types of *adaptive* mixture models [e.g., 152], where $s(x)$ are mixture weights in a simplex that have a strong inductive bias towards discretization. There are multiple ways to induce this bias towards discretization. For example, one can choose warping functions $s(x)$ which have sharp transitions between 0 and 1, such as the logistic sigmoid function. With multiple functions, $r \geq 2$, we can also explicitly penalize the warping functions from having similar values. Since each of these warping functions are constrained to be in $[0,1]$ this penalty would tend move their values towards 0 or 1. More generally, in the case of multiple functional regimes, we can penalize $s(x)$ from being far from $\{0,1\}$. For example, we could place a prior over $s(x)$ with a heavy weight on 1 and 0.

The flexibility of $s(x)$ defines the complexity of the change surface. In the simplest case, $x_i \in \mathbb{R}^1, s(x) \in \{0,1\}$, and the change surface reduces to a univariate changepoint used in much of the changepoint literature. Alternatively, if we consider $x \in \mathbb{R}^1, s(x) = \sigma(x)$ the change surface is a smooth univariate changepoint with a fixed rate of change. Such a model only permits a monotonic rate of change and single changepoint.

Comparison to changepoint models: We illustrate the difference between the warping functions, $s(x)$, of a change surface model and standard changepoint methods in Figure 2.2. The top plot shows unidimensional data with a clear change between two sinusoids. The subsequent plots represent the changes modeled in a discrete changepoint, sigmoid changepoint, and change surface model respectively. The changepoint model can only identify a change at a point in time, and the sigmoid changepoint is a special case of a change surface constrained to a fixed rate of change. However, a general change surface can model gradual changes as well as non-monotonic changes, providing a much richer representation of the data's dynamics, and seamlessly extending to multidimensional data.

Expressive change surfaces consider regimes as overlapping elements in the domain. They can illustrate if certain changes occur more slowly or quickly, vary over particular subpopulations, or change rapidly in certain regions of the input domain. Such insights are not provided by standard changepoint models but are critical for understanding policy interventions or scientific processes. Table 2.1 compares some of the limitations of changepoints with the added flexibility of change surfaces.

Yet the flexibility required by change surfaces as applied to real data sets might seem difficult to instantiate with any particular model. Indeed, machine learning methods are

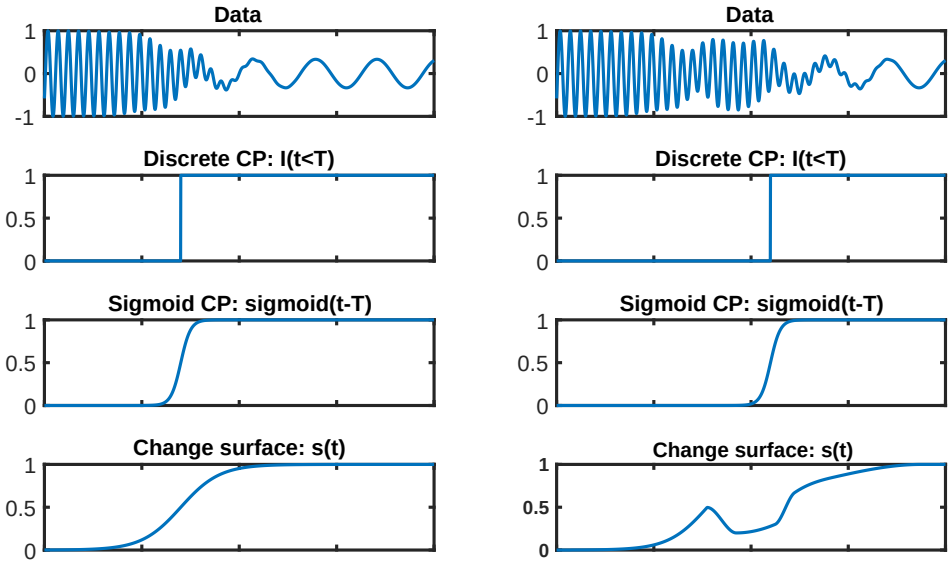


Fig. 2.2 Unidimensional comparison of changepoint and change surface methods. In each column, the top plot shows unidimensional data with a clear change between two sinusoids. The subsequent plots represent the warping functions of a discrete changepoint, sigmoid changepoint, and change surface model.

often desired to be expressive, interpretable, and scalable to large data. To address this challenge we introduce the Gaussian Process Change Surface (GPCS) in Section 2.3 which uses Gaussian process priors with flexible kernels to provide rich modeling capability, and a novel scalable inference scheme to permit the method to scale to massive data.

2.2.1 Change surface background model

In certain applications we are interested in modeling how a change occurs concurrent with a background function which is common to all regimes. For example, consider urban crime. If a police department staged a prolonged intervention in one sector of the city, we expect that some of the crime dynamics in that sector might change. However, seasonal and other weather-related patterns may remain the same throughout the entire city. In this case we want a model to identify and isolate those general background patterns as well as one or more clearly interpretable functions representing regions of change from the background distribution.

We can accommodate such a model as a special case of the generalized change surface from Eq. (2.2). Each latent function is modeled as $f_0(x) + f_i(x)$ where $f_0(x)$ models "back-

Table 2.1 Comparison of changepoint limitations to change surface flexibility.

Changepoints limited by:	**Change surfaces allow for:**
Considering unidimensional, often temporal-only problems	Multidimensional inputs with heterogeneous changes across the input dimensions. Indeed, we apply change surfaces to 3-dimensional, spatio-temporal problems in Section 2.4.
Detecting discrete or near-discrete changes in parameter distribution	Warping functions, $s(x)$, can be defined flexibly to allow for discrete or continuous changes with variable, and even non-monotonic rates of change.
Not simultaneously modeling the latent functional regimes	Learning $s_i(x)$ and $f_i(x)$ in Equation (2.1) to simultaneously model the change surface and underlying functional regimes.

ground" dynamics, and $f_i(x)$ models each *change* function. Since changes do not necessarily persist over the entire domain, we fix $f_r(x) = 0$, and allow $\sum_{i=1}^{r-1} s_i(x) \leq 1$. This approach results in the following *change surface background model*:

$$\begin{aligned} y(x) &= f_0(x) + s_1(x)f_1(x) + \cdots + s_{r-1}(x)f_{r-1}(x) + \varepsilon \\ &s.t. \\ &\sum_{i=1}^{r-1} s_i(x) \leq 1 \\ &s_i(x) \geq 0 \end{aligned} \tag{2.3}$$

Figure 2.3 presents a two-dimensional representation of the change surface and change surface background models. The data depicted comes from the numerical experiments in Section 2.4.1.

The explicit decomposition into background and change functions is valuable, for instance, if we wish to model *counterfactuals*: we want to know what the data in a region might look like had there been no change. The decomposition also enables us to interpret the precise effect of each change. Moreover, from a statistical perspective, the decomposition allows us to naturally encode inductive biases into the change surface model, allowing meaningful *a priori* statistical dependencies between each region. In the particular case of $r = 2$, the change surface background model has the form $y(x) = f_0(x) + s_1(x)f_1(x)$, where $f_1(x)$ is the only change function modulated by a change surface, $s_1(x) \in [0,1]$. This corresponds to observation studies or natural experiments where a single change is observed in the data. We explore this special case further in our discussion of counterfactual prediction, in Chapter 3.

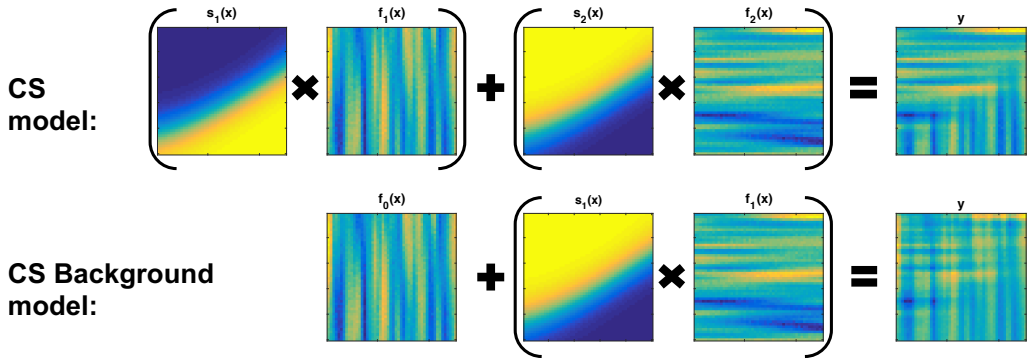


Fig. 2.3 Two-dimensional representation of the change surface model (Eq. 2.1) and change surface background model (Eq. 2.3).

Finally, for any change surface or change surface background model, it is critical that the model not overfit the data due to a proliferation of parameters, which could lead to erroneously detected changes even when no dynamic change is present. We discuss one strategy for preventing overfitting through the use of Gaussian processes in Section 2.3.

2.3 Gaussian Process Change Surfaces (GPCS)

We exemplify the general concept of change surfaces using Gaussian processes [e.g., 113]. We emphasize that our change surface formulations from Section 2.2 are not limited to a certain class of models. Yet Gaussian processes offer a compelling instantiation of change surfaces since they can flexibly model non-linear functions, seamlessly extend to multi-dimensional and irregularly sampled data, and provide naturally interpretable parameters. Perhaps most importantly, due to the Bayesian Occam's Razor principle [114, 94, 113, 153], Gaussian processes do not in general overfit the data, and extraneous model components are automatically pruned. Indeed, even though we develop a rich change surface model with multiple mixture parameters, our results below demonstrate that the model does not spuriously identify change surfaces in data.

Gaussian processes have been previously used for nonparametric changepoint modeling. Saatçi et al. [128] extend the sequential Bayesian Online Changepoint Detection algorithm [6] by using a Gaussian process to model temporal covariance within a particular regime. Similarly, Garnett et al. [51] provide Gaussian processes for sequential changepoint detection with mutually exclusive regimes. Moreover, Keshavarz et al. [82] prove asymptotic convergence bounds for a class of Gaussian process changepoint detection but are restricted to considering a single abrupt change in one-dimensional data. Focusing on anomaly detection,

Reece et al. [118] develop a non-stationary kernel that could conceivably be used to model a changepoint in covariance structure. However, as with most of the changepoint models discussed in Section 2.1, these models all focus on discrete changepoints, where regimes defined by distinct Gaussian processes change instantaneously.

A small collection of pioneering work has briefly considered the possibility of Gaussian processes with sigmoid changepoints [155, 92]. 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.

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 by modeling change surfaces with Gaussian processes. Thus our work both demonstrates a generalization of changepoint models and an enhancement to the expressive power of Gaussian processes.

2.3.1 Model specification

Change surface data consists of latent functions $f_1, \ldots, f_r$ defining r regimes in the data. The change surface defines the transitions between these functions. We could initially consider an input-dependent mixture model such as in Wilson et al. [152],

$$y(x) = w_1(x)f_1(x) + \cdots + w_r(x)f_r(x) + \varepsilon \tag{2.4}$$

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 $\sigma(z) : \mathbb{R}^r \to [0,1]^r$. These functions have support over the entire real line, but a range in $[0,1]$ and concentrated towards 0 and 1. Thus, each $w_i(x)$ in Eq. (2.4) becomes $\sigma_i(w(x))$, where $w(x) = [w_1(x), .., w_r(x)]$. Additionally, we choose $\sigma(z)$ such that it produces a convex combination over the weighting functions, $\sum_{i=1}^{r} \sigma_i(w(x)) = 1$. 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. Thus considering the general model of change surfaces in Eq. (2.1) we define each warping function as $s_i(x) = \sigma_i(w(x))$.

A natural choice for flexible change surfaces is to let $\sigma(z)$ be the softmax function. In this way the change surface can approximate a Heaviside step function, corresponding to the sharp transitions of standard changepoints, or more gradual changes. For r latent functions,

the resulting warping function is:

$$s_i(x) = \sigma_i(w(x)) = \text{softmax}(w(x))_i = \frac{\exp(w_i(x))}{\sum_{j=1}^{r} \exp(w_j(x))} \tag{2.5}$$

The Gaussian process change surface (GPCS) model is thus

$$y(x) = \sigma_1(w(x))f_1(x) + \cdots + \sigma_r(w(x))f_r(x) + \varepsilon \tag{2.6}$$

where each f_i is drawn from a Gaussian process. Importantly, we expect that each Gaussian process, $f_i(x)$, will have different hyperparameter values corresponding to different dynamics in the various regimes.

Since a sum of Gaussian processes is a Gaussian process, we can re-write Eq. (2.6) as $y(x) = f(x) + \varepsilon$, where $f(x)$ has a single Gaussian process prior with covariance function,

$$k(x,x') = \sigma_1(w(x))k_1(x,x')\sigma_1(w(x')) + \cdots + \sigma_r(w(x))k_r(x,x')\sigma_r(w(x')) \tag{2.7}$$

In this form we can see that $\sigma_1(w(x))\ldots\sigma_r(w(x))$ induce non-stationarity since they are dependent on the input x. Thus, even if we use stationary kernels for all k_i, GPCS observations follow a Gaussian process with a flexible, non-stationary kernel.

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,

$$w(x) = \beta_0 + \beta_1^T x \tag{2.8}$$

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 introduce a flexible $w(x)$ that is formed as a finite sum of Random Kitchen Sink (RKS) features which map the D dimensional input x to an m dimensional feature space. We use RKS features from a Fourier basis expansion with Gaussian parameters and employ marginal likelihood optimization to learn the parameters of this expansion. Similar expansions have been used to efficiently approximate flexible non-parametric Gaussian processes [88, 111].

Using m RKS features, $w(x)$ is defined as,

$$w(x) = \sum_{i=1}^{m} a_i \cos(\omega_i^T x_i + b_i) \tag{2.9}$$

where we initially sample,

$$a_i \sim \mathcal{N}(0, \frac{\sigma_0}{m} I) \tag{2.10}$$

$$\omega_i \sim \mathcal{N}(0, \frac{1}{4\pi^2} \Lambda^{-1}) \tag{2.11}$$

$$b_i \sim \text{Uniform}(0, 2\pi) \tag{2.12}$$

Initialization of hyperparameters σ_0 and diagonal matrix of length-scales, $\Lambda = \text{diag}(l_1^2, \ldots, l_D^2)$, is discussed in Section 2.3.3.

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. For example, in an industrial setting where we are modeling failure of parts in a factory we could define $w(x)$ such that it was monotonically increasing since machine parts do not self-repair. This bias could take the form of a linear function as in Equation (2.8). Note that since parameters are learned from data, the functional form of $w(x)$ does not require prior knowledge about if or where changes occur.

Kernel specification

Each latent function is specified by a kernel with its own set of 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 GPCS using multidimensional spectral mixture kernels [151] where $x \in \mathbb{R}^D$.

$$k_{\text{SM}}(x, x') = \sum_{q=1}^{Q} \omega_q \cos(2\pi(x - x')^T \mu_q) \prod_{d=1}^{D} \exp(-2\pi^2 (x^{(d)} - x'^{(d)})^2 v_q^{(d)}) \tag{2.13}$$

This kernel is derived via spectral densities that are scale-location mixtures of Q Gaussians. Each component in this mixture has mean $\mu_q \in \mathbb{R}^D$, covariance matrix $\text{diag}(v_q^{(1)}, ..., v_q^{(D)})$, and signal variance parameter $\omega_q \in \mathbb{R}^1$. 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 [153].

Previous work on Gaussian processes changepoint modeling has typically been restricted to RBF [128, 51] or exponential kernels [95]. However, expressive covariance functions are particularly critical for modelling multidimensional and spatio-temporal data – a key application for change surfaces – where structure is often complex and unknown a priori.

Initializing and training expressive kernels is often challenging. We propose a practical initialization procedure in Section 2.3.3, which can be used quite generally to help learn flexible kernels.

GPCS background model

Following Section 2.2.1 we extend GPCS to the "GPCS background model." For this model we add a latent background function, $f_0(x)$, with an independent Gaussian process prior. Using the same choices for expressive $w(x)$ and covariance functions, we define the GPCS background model as,

$$y(x) = f_0(x) + \sigma_1(w(x))f_1(x) + \cdots + \sigma_{r-1}(w(x))f_{r-1}(x) + \varepsilon \tag{2.14}$$

Recall that in this model we set $f_r(x) = 0$. Additionally, since we continue to enforce $\sum_{i=1}^{r} \sigma_i(w(x)) = 1$, thus $\sum_{i=1}^{r-1} \sigma_i(w(x)) \leq 1$.

This model effectively places different priors on the background and change regions, as opposed to the the standard GPCS model which places the same GP prior on each regime. The different priors in the GPCS background model reflect an intentional inductive bias which could be advantageous in certain domain settings, such as policy interventions, as discussed in Section 2.2.1 above.

2.3.2 Scalable inference

Analytic optimization and inference for Gaussian processes requires computation of the log marginal likelihood from Eq. (1.9). Yet solving linear systems and computing log determinants over $n \times n$ covariance matrices, using standard approaches such as the Cholesky decomposition, requires $O(n^3)$ computations and $O(n^2)$ memory, which is impractical for large datasets. Recent advances in scalable Gaussian processes [155] have reduced this computational burden by exploiting Kronecker structure under two assumptions: (1) the inputs lie on a grid formed by a Cartesian product, $x \in X = X^{(1)} \times ... \times X^{(D)}$; and, (2) the kernel is multiplicative across each dimension. Multiplicative kernels are commonly

employed in spatio-temporal Gaussian process modeling [96, 95, 50], corresponding to a soft a priori assumption of independence across input dimensions, without ruling out posterior correlations. The popular RBF and ARD kernels, for instance, already have this multiplicative structure. 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 [127] shows how one can solve linear systems and compute log determinants in $O(Dn^{\frac{D+1}{D}})$ operations using $O(Dn^{\frac{2}{D}})$ memory. Furthermore, Wilson et al. [153] extends the Kronecker methods for incomplete grids. Yet for additive compositions of kernels, such as those needed for change surface modeling in Eq. (2.7), the resulting sum of matrix Kronecker products does not decompose as a Kronecker product. Thus, the standard Kronecker approaches for scalable inference and learning are inapplicable. Instead, solving linear systems for the kernel inverse can be efficiently carried out through linear conjugate gradients as in Flaxman et al. [50] that only rely on matrix vector multiplications, which can be performed efficiently with sums of Kronecker matrices.

However, there is no exact method for efficient computation of the log determinant of the sum of Kronecker products. Instead, Flaxman et al. [50] upper bound the log determinant using the Fiedler bound [49] 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,

$$\log(|A+B|) \leq \sum_{i=1}^{n} \log(\alpha_i + \beta_{n-i+1}) \tag{2.15}$$

While efficient, the Fiedler bound does not generalize to more than two matrices.

Weyl bound

In order to achieve scalable computations for an arbitrary additive composition of Kronecker matrices, we propose to bound the log determinant of the sum of multiple covariance matrices using Weyl's inequality [150] 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,

$$\mu_{i+j-1} \leq \alpha_i + \beta_j \quad \forall i, j \geq 1 \tag{2.16}$$

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 use for bounding μ_{i+j-1}. Thus for each eigenvalue, μ_k, we wish to choose i, j that

minimizes $\alpha_i + \beta_j$ subject to $k = i + j - 1$. One might be tempted to minimize over all possible pairs for each eigenvalue, $\mu_1, \ldots, \mu_n$, in order to obtain the tightest bound on the log determinant. Unfortunately, such a procedure requires $O(n^2)$ computations. Instead we explore two possible alternatives:

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

2. We employ a greedy search to choose the minimum of v possible pairs of eigenvalues. Using the previous i' and j', we consider $\{\alpha_i, \beta_j\}$ for all $i = i' - \frac{v}{2}, ..., i' + \frac{v}{2}$ and the corresponding j values. Setting $v = 1$ corresponds to the middle heuristic. Setting $v = n$ corresponds to the exact Weyl bound. The greedy search requires $O(vn)$ computations.

In addition to bounding the sum of kernels, we must also deal with the scaling functions, $\sigma_i(w(x))$. We can rewrite Eq. (2.7) in matrix notation,

$$K = S_1 K_1 S_1' + \cdots + S_r K_r S_r' \tag{2.17}$$

where $S_i = \text{diag}(\sigma_i(w(x)))$ and $S_i' = \text{diag}(\sigma_i(w(x')))$. Employing the bound on eigenvalues of matrix products [21],

$$\text{sort}(\text{eig}(AB)) \leq \text{sort}(\text{eig}(A))\text{sort}(\text{eig}(B)) \tag{2.18}$$

we can bound the log determinant of K in Eq. (2.17) with an iterative Weyl approximation over $[\{s_{i,l} k_{i,l} s'_{i,l}\}_{l=1}^{n}]_{i=1}^{r}$ where $s_{i,l}$, $k_{i,l}$, and $s'_{i,l}$ are the l^{th} largest eigenvalue of S_i, K_i, and S_i' respectively.

We empirically evaluate the exact Weyl bound, middle heuristic, and greedy search with $v = 80$ pairs of eigenvalue indexes to search above and below the previous index. All experiments are evaluated using GPCS with synthetic data generated according to the procedure in Section 2.4.1. We also compare these results against the Fiedler bound in the case of two kernels.

Figure 2.4 depicts the ratio of each approximation to the true log determinant, and the time to compute each approximation over increasing number of observations for two kernels. While the Fiedler approximation is more accurate than any Weyl approach, all approximations perform quite similarly (note the fine grained axis scale) and converge to ≈ 0.85 of the true log determinant. In terms of computation time, the exact Weyl bound scales poorly with data size as expected. Yet both approximate Weyl bounds scale well. In practice, we use the middle heuristic described above, since it provides the fastest results, nearly equivalent to the Fiedler bound.

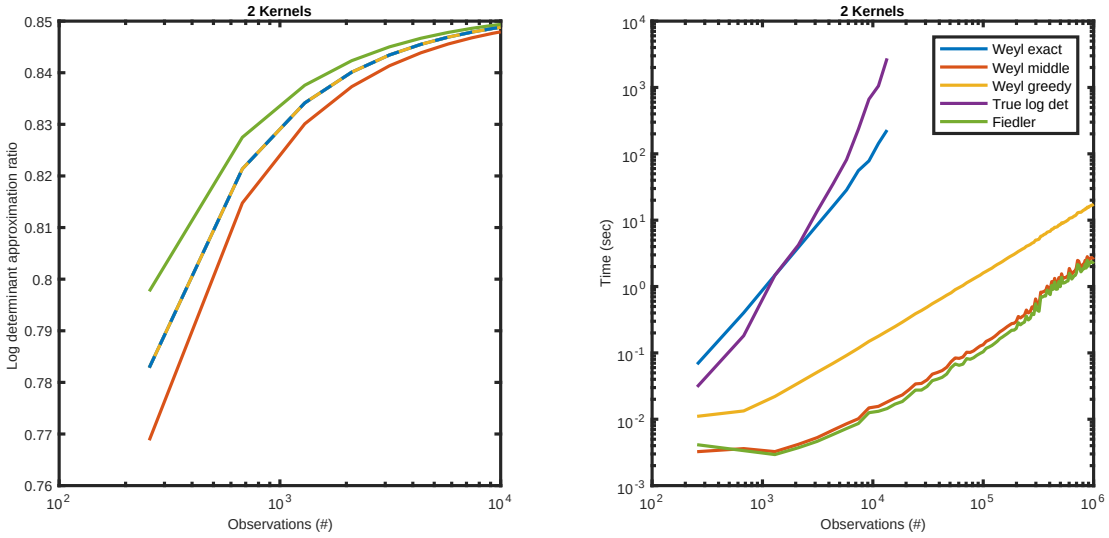


Fig. 2.4 Left plot shows the ratio of log determinant approximations to the true log determinant of two additive kernels. Note that the y-axis is scaled to a relatively narrow band. The dashed line indicates that both the Weyl exact and Weyl greedy method performed similarly. Right plot shows the time to compute each approximation and the true log determinant.

Figure 2.5 depicts the same quantities as Figure 2.4 but using three additive kernels. Since the Fiedler approximation is only valid for two kernels it is excluded from these plots. While the log determinant approximation ratios are less accurate for small datasets, as the data size increases all Weyl approximations converge to ≈ 0.8.

In addition to enabling scalable change surface kernels, the Weyl bound method permits scalable additive kernels in general. When applied to the spatio-temporal domain this yields the first scalable Gaussian process model which is non-separable in space and time.

Massively Scalable Inference

We further extend the scalability and flexibility of the Weyl bound method by leveraging a structured kernel interpolation methodology from the KISS-GP framework [154]. Although many spatiotemporal policy relevant applications naturally have near-grid structure, such as readings over a nearly dense set of latitudes, longitudes, and times, this integration with KISS-GP further relaxes the dependencies on grid assumptions. The resulting approach scales to much larger problems by interpolating data to a smaller, user-defined grid. In particular, with local cubic interpolation, the error in the kernel approximation is upper bounded $O(1/m^3)$ for m latent grid points, and m can be very large because the kernel matrices in this space are structured. These scalable approaches are thus very generally applicable as demonstrated

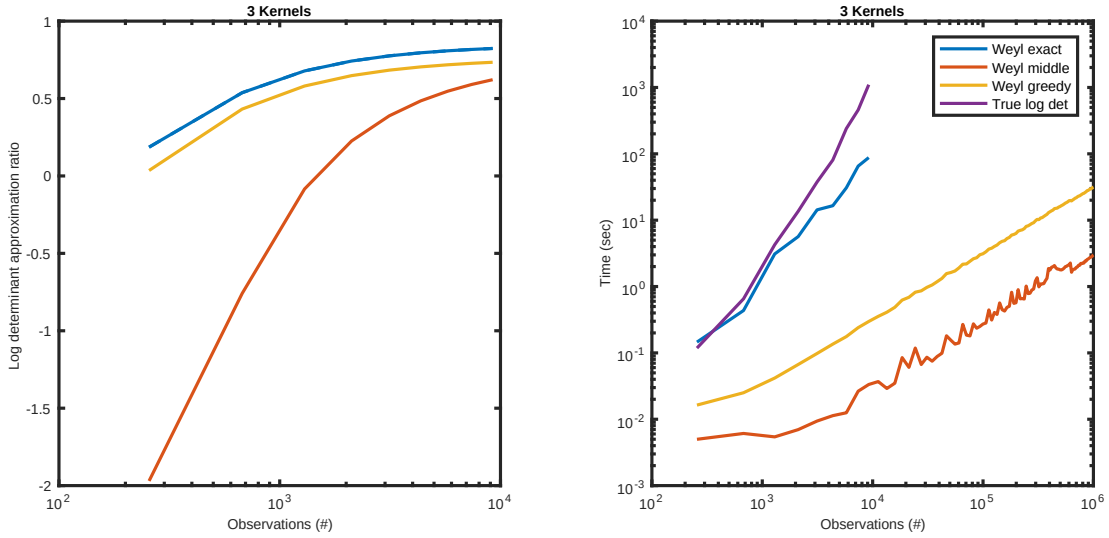


Fig. 2.5 Left plot shows the ratio of approximations to the true log determinant of 3 additive kernels. Note that the y-axis has a much larger scale than in Figure 2.4. Right plot shows the time to compute each approximation and the true log determinant of 3 additive kernels.

in an extensive range of previously published experiments in [156, 157] based on these techniques. Additionally, KISS-GP enables the Weyl bound approximation methods to apply to arbitrary, non-grid data.

We empirically demonstrate the advantages of integration with KISS-GP by evaluating an additive GPCS on the two-dimensional data described above. Although the original data lies on a grid, we use KISS-GP interpolation to compute the negative log likelihood on four grids of increasingly smaller size. Figure 2.6 depicts the negative log likelihood and the computation time for these experiments using the Weyl middle heuristic. The plot legend indicates the size of the induced grid size. For example, 'KISS-GP 75%' is 75% the size of the original grid. Note that the time and log likelihood scales in Figure 2.6 are different from those in Figures 2.4 and 2.5 since we are now computing full inference steps as opposed to just computing the log determinant. The results indicate that with minimal error in negative log likelihood accuracy we can substantially reduce the time for inference.

2.3.3 Initialization

Since GPCS uses flexible spectral mixture kernels, as well as RKS features for the change surface, the parameter space is highly multimodal. Therefore, it is essential to initialize the model hyperparameters appropriately. Below we present an approach where we first initialize the $w(x)$ RKS features and then use those values in a novel initialization method for

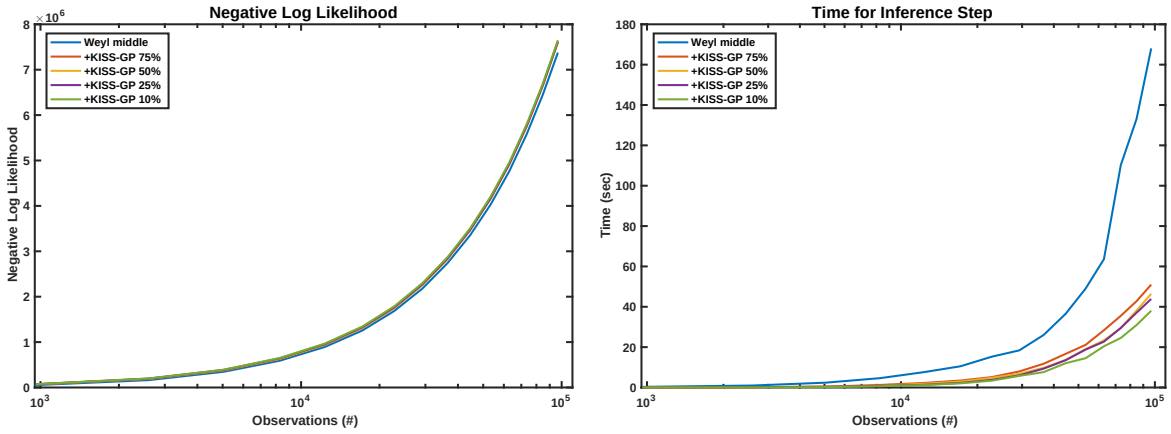


Fig. 2.6 Plots showing negative log likelihood and time for inference on two additive kernels using the Weyl bound on grids of decreasing size. For example, 'KISS-GP 75%' computes the Weyl middle bound on a grid which is 75% the size of the original grid used to compute the first line.

the spectral mixture kernels. Like most GP optimization problems, GPCS hyperparameter optimization is non-convex and there are no provable guarantees that the proposed initialization will result in optimal solutions. However, it is our experience that this initialization procedure works well in practice for the GPCS as well as spectral mixture kernels in general.

To initialize $w(x)$ defined by RKS features we first simplify the change surface model by assuming that each latent function, $f_1, ..., f_r$, from Eq. (2.6) is drawn from a Gaussian process with an RBF kernel. Since RBF kernels have far fewer hyperparameters than spectral mixture kernels, starting with RBF kernels helps our approach find good starting val

ues for $w(x)$. 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 as a terminal model.

Algorithm 1 Initialize RKS $w(x)$ by optimizing a simplified model with RBF kernels

1: **for** $i = 1 : m_1$ **do**
2: Draw a, ω, b for RKS features in $w(x)$
3: Draw m_2 sets of hyperparameter values for RBF kernels, $\{\theta_1, ..., \theta_{m_2}\}$
4: Choose the best hyperparameter set, $\theta^{(i)} = \texttt{max-likelihood}(\theta_1, ..., \theta_{m_2})$
5: Partial optimization of $\{a, \omega, b, \theta\} \rightarrow \Theta^{(i)}$
6: **end for**
7: Choose the best set of hyperparameters, $\Theta = \texttt{max-likelihood}(\Theta^{(1)}, ..., \Theta^{(m_1)})$
8: Optimize Θ 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 2.3.1) m_1 number of

times. We set reasonable values for hyperparameters in those prior distributions. Specifically, we let $\Lambda = (\frac{\text{range}(x)}{2})^2$, $\sigma_0 = \text{std}(y)$, and $\sigma_n = \frac{\text{mean}(|y|)}{10}$. These choices are similar to those employed in [88].

For each sampled set of $w(x)$ hyperparameters, we sample m_2 sets of hyperparameters for the RBF kernels and select the set with the highest marginal likelihood. Then we run an abbreviated optimization procedure over the combined $w(x)$ and RBF hyperparameters and select the joint set that achieves the highest marginal likelihood. Finally, we optimize the resulting hyperparameters until convergence.

In order to initialize the spectral mixture kernels, we use the initialized $w(x)$ from above to define the subset $\{x : \sigma_i(w(x)) > 0.5\}$ where each latent function, f_i from Eq. (2.6), is dominant. We then take a Fourier transform of $y(x)$ over each dimension, $x^{(d)}$, of $\{x : \sigma_i(w(x)) > 0.5\}$ 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 [155]. Since spectral mixture kernels model the spectral density with Q Gaussians on $\mathbb{R}^1$, we fit a 1-dimensional Gaussian mixture model,

$$p(x) = \sum_{q=1}^{Q} \phi_q \mathcal{N}(\mu_q, v_q) \tag{2.19}$$

to the empirical spectrum for each dimension. Using the learned mixture model we initialize the parameters of the 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 $x^{(d)} \in \{x : \sigma_i(w(x)) > 0.5\}$
    Sample $s \sim |\text{FFT}(\text{sort}(y(x^{(d)})))|^2$
    Fit Q component GMM as $p(s) = \sum_{q=1}^{Q} \phi_q^{(d)} \mathcal{N}(\mu_q^{(d)}, v_q^{(d)})$
    Initialize $\omega_q = \text{std}(y(x^{(d)})) * \phi_q$
  end for
end for

After initializing $w(x)$ and spectral mixture hyperparameters, we jointly optimize the entire model using marginal likelihood and non-linear conjugate gradients [115].

2.4 Experiments

We demonstrate the power and flexibility of GPCS by applying the model to a variety of numerical simulations and complex human settings. We begin with 2-dimensional numerical data in Section 2.4.1, and show that GPCS is able to correctly model out-of-class polynomial change surfaces, and that it provides higher accuracy regressions than other comparable methods.

We next consider coal mining, epidemiological, and urban policy data to provide additional analytical evidence for the effectiveness of GPCS and to demonstrate how GPCS results can be used to provide novel policy-relevant and scientifically-relevant insights. The ground truth against which GPCS is evaluated are the domain specific interventions in these case studies.

In order to compare GPCS to standard changepoint models, we use a 1-dimensional dataset on the frequency of coal mining accidents. After fitting GPCS, we show that the change surface is able to identify a region of change similar to other changepoint methods. However, unlike changepoint methods that only identify a single moment of change, GPCS models how the data changes over time.

We then employ GPCS to analyze two complex spatio-temporal settings involving policy and scientific questions. First we examine requests for residential lead testing kits in New York City between 2014-2016, during a time of heightened concern about lead-tainted water. GPCS identifies a spatially and temporally varying change surface around the period when issues of water contamination were being raised in the news. We conduct a regression analysis on the resulting change surface features to better understand demographic factors that may have affected residents' concerns about lead-tainted water.

Second, we apply GPCS to model state-level measles incidence in the United States during the twentieth century. GPCS identifies a substantial change around the introduction of the measles vaccine in 1963. However, the shape of the change surface varies over time for each state, indicating possible spatio-temporal heterogeneity in the adoption and effectiveness of the vaccination program during its initial years. We use regression analysis on the change surface features to explore possible institutional and demographic factors that may have influenced the impacts of the measles vaccination program.

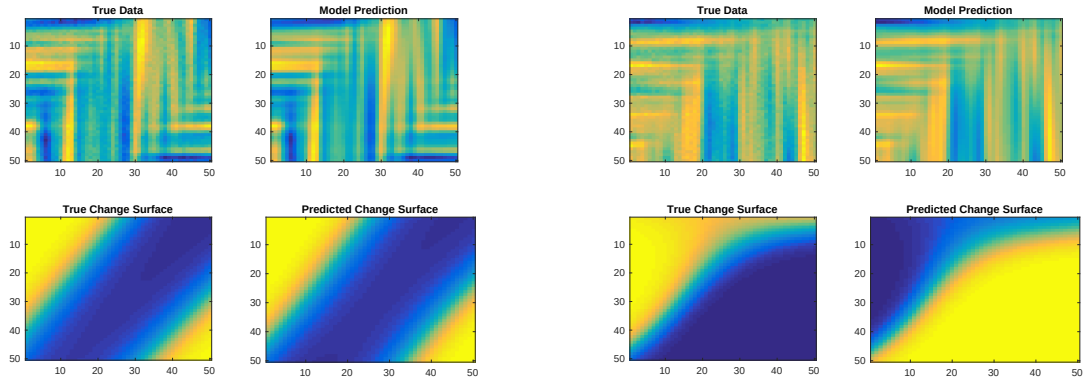
2.4.1 Numerical Experiments

We generate a 50×50 grid of synthetic data by drawing independently from two latent functions, $f_1(x)$ and $f_2(x)$. Each function is characterized by an independent Gaussian process with a two-dimensional RBF kernel of different length-scales and signal vari-

ances. The synthetic change surface between the functions is defined by $\sigma(w_{\text{poly}}(x))$ where $w_{\text{poly}}(x) = \sum_{i=0}^{3} \beta_i^T x^i$, $\beta_i \sim \mathcal{N}(0, 3I_D)$. We chose a polynomial change surface because it generates complex change patterns but is out-of-class when we use RKS features for $w(x)$, thus testing the robustness of GPCS to change surface misspecification.

GPCS model

Using the synthetic data generation technique described above we simulate data as $y = \sigma(w_{poly}(x))f_1(x) + (1 - \sigma(w_{\text{poly}}(x)))f_2(x) + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. We apply GPCS with two latent functions, spectral mixture kernels, and $w(x)$ defined by RKS features. We do not provide the model with prior information about the change surface or latent functions. As emphasized in Section 2.3.3, successful convergence is dependent on reasonable initialization. Therefore, we use $m_1 = 100$ and $m_2 = 20$ for Algorithm 1. Figure 2.7 depicts two typical results using the initialization procedure followed by analytic optimization. The model captures the change surface and produces an appropriate regression over the data. Note that in Figure 2.7b the predicted change surface is flipped since the order of functions is not important in GPCS.



(a) (b)

Fig. 2.7 Two numerical data experiments. In each of (a) and (b) the top-left plot depicts the data (e.g., observations indexed by two dimensional spatial inputs); the bottom-left shows the true change surface with the range from blue to yellow depicting $\sigma_1(w(x))$. The top-right depicts the predicted output; the bottom-right shows the predicted change surface. Note that the predicted change surface in plot (b) is flipped since the order of functions is not important.

To demonstrate that the initialization method from Section 2.3.3 provides consistent results, we consider a numeric example and run GPCS 30 times with different random seeds. Figure 2.8 provides the true data and change surface as well as the mean and standard

deviation over the 30 experimental results using the Section 2.3.3 initialization procedure. For the predicted change surface we manually normalized the orientation of the change surface before computing summary statistics. The results illustrate that the initialization procedure provides accurate and consistent results for both y and $\sigma(w(x))$ across multiple runs. Indeed, when we repeat these experiments with random initialization, instead of the procedure from Section 2.3.3, the MSE between the predicted and true change surface is 58% greater than when using our initialization procedure. Additionally, the results have a 17% larger standard deviation of $\sigma(w(x_i))$ over the 30 runs, demonstrating that the produre we propose provides more consistent and accurate results.

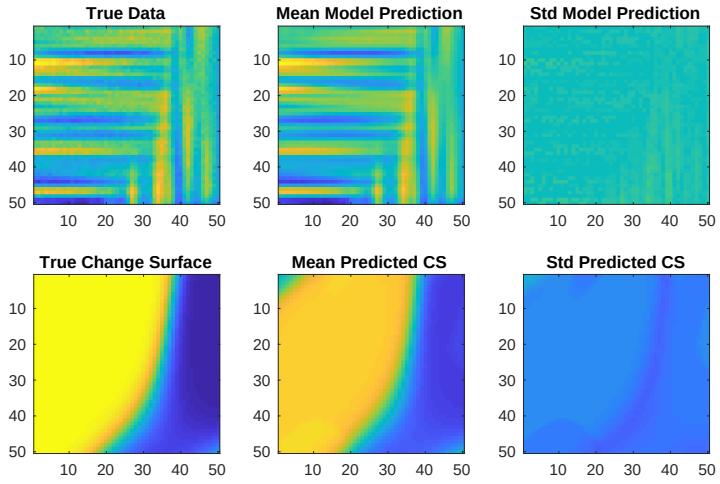


Fig. 2.8 Consistency results across 30 runs with different random seeds. True data and change surface are on the left, while the mean and standard deviation of the predicted results are in center and right panels.

Using synthetic data, we create a predictive test by splitting the data into training and testing sets. We compare GPCS to three other expressive, scalable methods: sparse spectrum Gaussian process with 500 basis functions [88], sparse spectrum Gaussian process with fixed spectral points with 500 basis functions [88], and a Gaussian process with multiplicative spectral mixture kernels in each dimension. For each method we average the results for 10 random restarts. For each method Table 2.2 shows the normalized mean squared error (NMSE),

$$\text{NMSE} = \frac{\|y_{test} - y_{pred}\|_2^2}{\|y_{test} - \bar{y}_{train}\|_2^2} \tag{2.20}$$

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

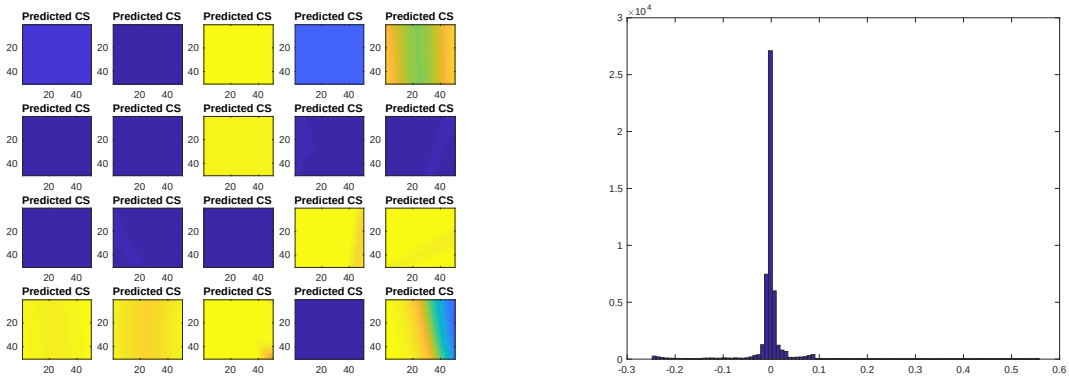
Table 2.2 Comparison of prediction accuracy (normalized mean squared error) using flexible and scalable Gaussian process methods on synthetic multidimensional change-surface data.

Method	**NMSE**
GPCS	0.00078
SSGP	0.01530
SSGP fixed	0.02820
Spectral mixture	0.00200

GPCS performed best due to the expressive non-stationary covariance function that fits to the different functional regimes in the data. Although the other methods can flexibly adapt to the data, they must account for the change in covariance structure by setting a shorter global length-scale over the data, thus underestimating the correlation of points in each regime. Thus their predictive accuracy is lower than GPCS, which can accommodate changes in covariance structure across the boundary of a change surface while retaining reasonable covariance length-scales within each regime.

As discussed in Section 2.3, the underlying probabilistic Gaussian process model behind GPCS automatically *discourages* extraneous complexity, favoring the simplest explanations consistent with the data [94, 114, 113, 153, 155]. This property enables GPCS to discover interpretable generative hypothesis for the data, which is crucial for public policy applications. This Bayesian Occam's razor principle is a cornerstone of many probabilistic approaches, such as automatically determining the intrinsic dimensionality for probabilistic PCA [98], or hypothesis testing for Bayesian neural network architectures [94]. In the absence of such automatic complexity control, these methods would always favor the highest intrinsic dimensionality or the largest neural network respectively.

To demonstrate this Occam's razor principle in our context, we generate numeric data from a single GP without any change surface by setting $\sigma(w_{\text{poly}}(x)) = 0$, and fit a *misspecified* GPCS model assuming two latent regimes. Figure 2.9 depicts the predicted change surfaces for 20 experiments of such data. The left panel illustrates pictorially that the change surfaces are nearly all flat at either $\sigma_1(w(x)) = 0$ or $\sigma_1(w(x)) = 1$ for these experiments. Specifically, $\text{std}[\sigma_1(w(x))] < 0.03$ for all but two runs. This finding indicates that GPCS discovers that no dynamic transition exists and does not overfit the data, despite the added flexibility afforded by multiple mixture components. Only one of the 20 results (bottom-right) indicates a change, and even in that case the magnitude of the transition is markedly subdued as compared to the results in Figures 2.7 and 2.10. While the upper-right result appears to have a large transition, in fact it has a flat change surface with $\text{std}[\sigma_1(w(x))] = 0.07$. The right panel provides a histogram of the mean centered change surface values for all experiments,



(a) (b)

Fig. 2.9 Data without any change surface, $\sigma(w_{\text{poly}}(x)) = 0$. The left panel depicts $\sigma_1(w(x))$ for each experiment. The right panel provides a histogram of the mean centered change surfaces values, $\sigma_1(w(x)) - \sum_{i \in n} \sigma_1(w(x_i))$.

$\sigma_1(w(x)) - \sum_{i \in n} \sigma_1(\omega(x_i))$, again demonstrating that GPCS learns very flat change surfaces and does not erroneously identify a change.

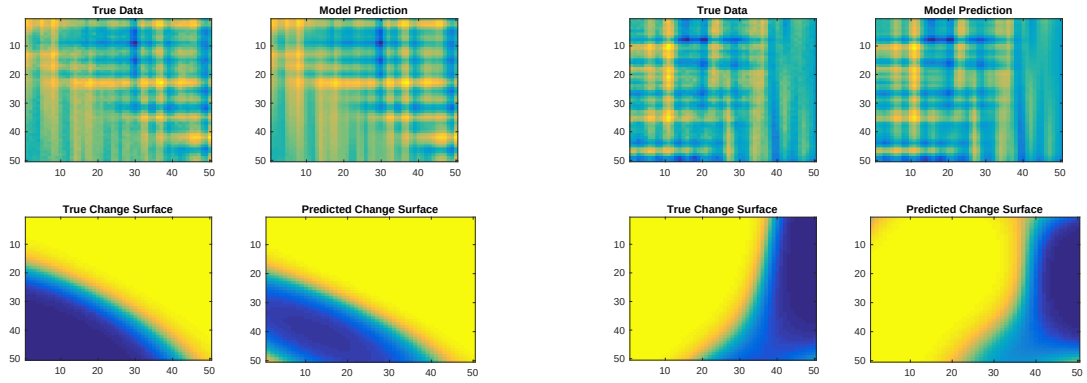
GPCS background model

We test the GPCS background model with a similar setup. Using the synthetic data generation technique described above, we simulate data as $y = f_0(x) + \sigma(w_{\text{poly}}(x)) f_1(x) + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. We again note that the polynomial change surface is out-of-class.

We apply the GPCS background model with one background function and one latent function scaled by a change surface. Both Gaussian process priors use spectral mixture kernels, and $w(x)$ is defined by RKS features. We do not provide the model with prior information about the change surface or latent functions. Figure 2.10 depicts two typical results using the initialization procedure followed by analytic optimization. The model captures the change surface and produces an appropriate regression over the data.

Log Gaussian Cox Process

The numerical experiments above demonstrate the consistency of GPCS for identifying out-of-sample change surfaces and modeling complex data for high accuracy prediction. To further demonstrate the flexibility of the model, we apply GPCS to data generated by a log-Gaussian Cox process [99, 50]. This inhomogeneous Poisson process is modulated by a



(a) (b)

Fig. 2.10 Two numerical data experiments. In each of (a) and (b) the top-left plot depicts the data; the bottom-left shows the true change surface with the range from blue to yellow depicting $\sigma_1(w(x))$. The top-right depicts the predicted output; the bottom-right shows the predicted change surface.

stochastic intensity defined as a GP,

$$\lambda = f \tag{2.21}$$

$$f \sim \mathscr{GP}(\mu, K) \tag{2.22}$$

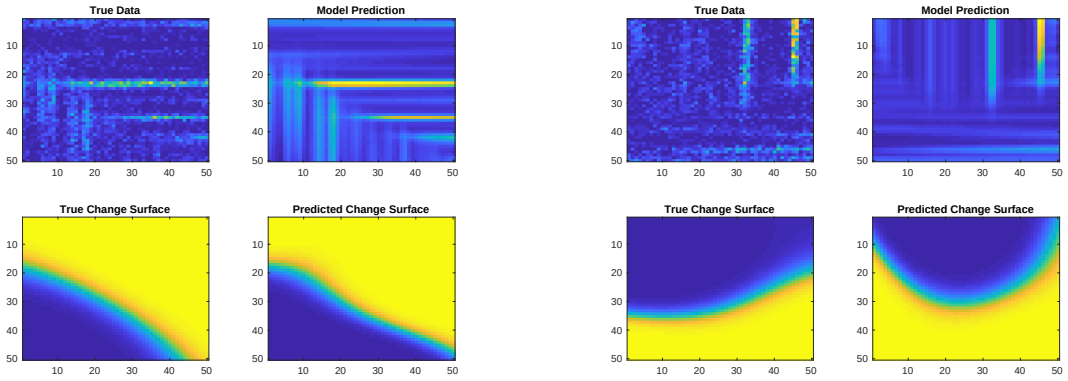
Conditional on λ, and letting s denote a region in space-time, the resulting small-area count data are non-negative integers distributed as

$$y(s) \,|\, \lambda \sim \text{Poisson}\big(\exp \int_s \lambda(x)dx\big). \tag{2.23}$$

We let this GP model be a convex combination of two GPs with an out-of-sample change surface, as described in Section 2.4.1. Thus we generated data from this model as

$$y \,|\, f_1(x_i), f_2(x_i) \sim \text{Poisson}\Big(\exp\big[\sigma(w_{poly}(x))f_1(x) + (1-\sigma(w_{\text{poly}}(x)))f_2(x) + \varepsilon\big]\Big). \tag{2.24}$$

Such data substantially departs from the type of data that GPCS is designed to model. Indeed, while custom approaches are often created to handle inhomogeneous Poisson data [50, 137], we use GPCS to demonstrate its flexibility and applicability to complex non-Gaussian data. The results are shown in Figure 2.11. The model provides accurate change surfaces and predictions even though the data is substantially out-of-class – even beyond the out-of-class change surface data from Sections 2.4.1 and 2.4.1. The precise location of change surfaces



(a) (b)

Fig. 2.11 Two numerical data experiments with data from a log-Gaussian Cox process. In each of (a) and (b) the top-left plot depicts the data (e.g., observations indexed by two dimensional spatial inputs); the bottom-left shows the true change surface with the range from blue to yellow depicting $\sigma_1(w(x))$. The top-right depicts the predicted output; the bottom-right shows the predicted change surface.

deviates slightly in GPCS, particularly on the left edge of Figure 2.11b where the raw data is highly stochastic. Additionally, the model predictions are smoothed versions of the true latent data, which reflects the fundamental difference between Gaussian and Poisson models.

2.4.2 British Coal Mining Data

British coal mining accidents from 1861 to 1962 have been well studied as a benchmark in the point process and changepoint literature [110, 26, 6]. We use yearly counts of accidents from Jarrett [77]. Adams and MacKay [6] indicate 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 [133]. We apply GPCS to show that it can detect changes corresponding to policy interventions in data while providing additional information beyond previous changepoint approaches.

We consider GPCS with two latent functions, spectral mixture kernels, and $w(x)$ defined by RKS features. We do not provide the model with prior information about the 1887 legislation date. Figure 2.12 depicts the cumulative data and predicted change surface. The red line marks the year 1887 and the magenta line marks $x : \sigma(w(x)) = 0.5$. GPCS correctly identified the change region and suggests a gradual change that took 5.6 years to transition from $\sigma(w(x)) = 0.25$ to $\sigma(w(x)) = 0.75$.

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)

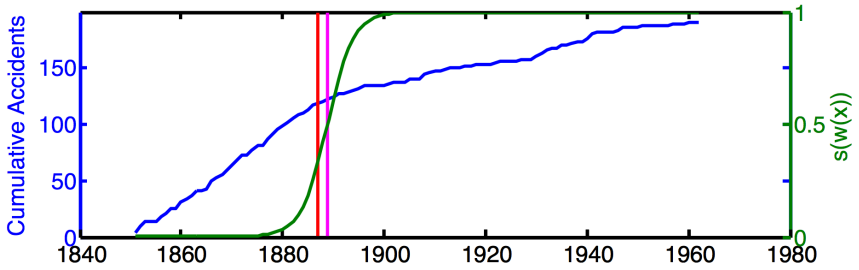


Fig. 2.12 British coal mining accidents from 1851 to 1962. The blue line depicts cumulative annual accidents, the green line plots $\sigma(w(x))$, the vertical red line marks the Coal Mines Regulation Act of 1887, and the vertical magenta line indicates $\sigma(w(x)) = 0.5$.

Table 2.3 Comparing methods for estimating the date of change in coal mining data.

Method	**Estimated date**
GPCS $\sigma(w(x)) = 0.5$	1888.8
PELT mean change	1886.5
PELT variance change	1882.5
ecp	1887.0
Student-t test	1886.5
Bartlett test	1947.5
Mann-Whitney test	1891.5
Kolmogorov-Smirnov test	1896.5

[83] for changes in mean and variance and a nonparametric method named "ecp" [76]. Additionally, we tested the batch changepoint method described in [120] with Student-t and Bartlett tests for Gaussian data as well as Mann-Whitney and Kolmogorov-Smirnov tests for nonparametric changepoint estimation [134]. Figure 2.3 compares the dates of change identified by these methods to the midpoint date where $\sigma(w(x)) = 0.5$ in GPCS.

Most of the methods identified a midpoint date between 1886 and 1895. While each method provides a point estimate of the change, only GPCS provides a clear, quantitative description of the development of this change. Indeed the 5.6 years during which the change surface transitions between $\sigma(w(x)) = 0.25$ to $\sigma(w(x)) = 0.75$ nicely encapsulate most of the point estimate method results.

2.4.3 New York City Lead Data

In recent years there has been heightened concern about lead-tainted water in major United States metropolitan areas. For example, concerns about lead poisoning in Flint, Michigan's water supply garnered national attention in 2015 and 2016, leading to Congressional hearings.

Similar lead contamination issues have been reported in a spate of United States cities such as Cleveland, OH, New York, NY, and Newark, NJ [47]. 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 [52].

In order to understand the evolving dynamics of New York City residents' concerns 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 [36]. 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 2.13 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 [30].

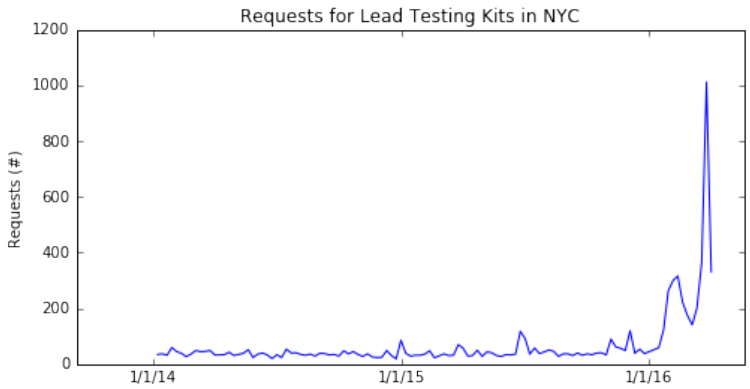


Fig. 2.13 Requests for residential lead testing kits in New York City aggregated at a weekly level across the entire city.

While there is a distinct uptick in requests for kits towards the middle and end of the observation period, there is no ground truth change point, unlike the coal mining example in Section 2.4.2 and the measles incidence example in Section 2.4.4. We apply GPCS with two latent functions, spectral mixture kernels, and $w(x)$ defined by RKS features. Note that the inputs are three dimensional, $x \in \mathbb{R}^3$, with two spatial dimensions representing centroids of each zipcode and one temporal dimension.

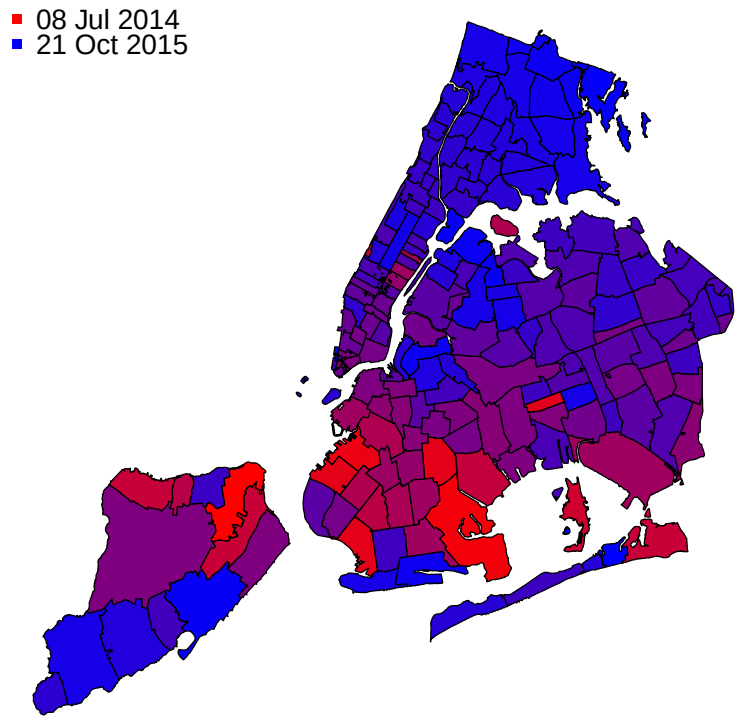


Fig. 2.14 NYC zip codes colored by the date where $\sigma(w(x_{\text{zip}})) = 0.5$. Red indicates earlier dates, with Bulls Head in Staten Island being the earliest. Blue indicates later dates, with New Hyde Park at the eastern edge of Queens being the latest.

The model suggests that residents' concerns about lead tainted water had distinct spatial and temporal variation. In Figure 2.14 we depict the midpoint, $\sigma(w(x_{\text{zip}})) = 0.5$, for each zip code. We illustrate the spatial variation in the midpoint date by shading zip codes with an early midpoint in red and zip codes with later midpoint in blue. Regions in Staten Island and Brooklyn experienced the earliest midpoints, with Bulls Head in Staten Island (zip code 10314) being the first area to reach $\sigma(w(x_{\text{zip}})) = 0.5$ 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 surrounding the Flint water crisis. However, most zip codes have midpoint dates sometime in 2015.

In Figure 2.15 we depict the change surface slope from $\sigma(w(x_{\text{zip}})) = 0.25$ to $\sigma(w(x_{\text{zip}})) = 0.75$ for each zip code to estimate the rate of change. We illustrate the variation in slope by shading zip codes with flatter change slopes in red and the steeper change slopes 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.

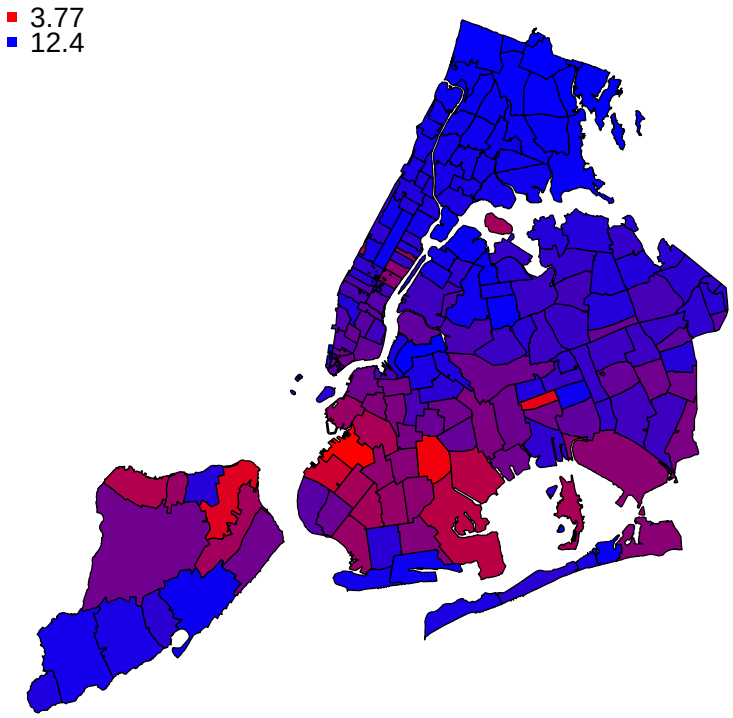


Fig. 2.15 NYC zip codes colored by the slope of $\sigma(w(x_{\text{zip}}))$ from 0.25 to 0.75. Red indicates flatter slopes, with Mariner's Harbor in Staten Island being the flattest. Blue indicates steeper slopes, with Woodlawn Heights in the Bronx being the steepest.

Regression analysis: The 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 midpoint date between the two regimes. All data were taken from the 2014 American Community Survey 5 year average at the zip code level [31]. Factors considered included information about residents such as 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, we considered information about when the homes were built.

Results of a linear regression over all factors can be seen in Table 2.4. Five variables were statistically significant at a p-value < 0.05: median annual household income, percentage of houses built 1940-1959, percentage of householders with high school equivalent education, percentage of householders with at least a college education, and percentage of owner occupied households. Median annual household income had a positive correlation with the

Table 2.4 Results from a linear regression to the NYC lead midpoint date, $\sigma(w(x_{\text{zip}})) = 0.5$. Variables are listed on the left while their coefficients, with standard errors in parentheses, are listed on the right. Asterisks indicate statistically significant variables.

	Dependent variable:
	Midpoint date
Log median household income	21.916**
	(7.912)
% homes built after 2010	0.549
	(0.724)
% homes built 2000-2009	0.061
	(0.164)
% homes built 1980-1999	−0.070
	(0.153)
% homes built 1960-1979	0.027
	(0.094)
% homes built 1940-1959	0.667**
	(0.092)
% education high school equivalent	−1.609**
	(0.331)
% education some college	0.143
	(0.312)
% education college and above	−0.864**
	(0.303)
% households owner occupied	−0.310*
	(0.126)
Average family size	9.507
	(6.453)
% households with member 18 or younger	−0.020
	(0.282)
% households with member 60 or older	0.202
	(0.215)
% households with only one member	0.283
	(0.227)
Constant	−149.602
	(77.036)
Observations	176
R^2	0.420
Adjusted R^2	0.370
Note:	*p<0.05; **p<0.01

change date, suggesting that higher household income is associated with later midpoint dates. People with lower incomes may tend to live in housing that is less well maintained, or is perceived to be less well maintained. Thus they may require less "activation energy" to request lead testing kits when faced with possible environmental hazards. Education levels were compared to a base value of householders with less than a high school education. Thus zip codes with more educated householders tended to have earlier midpoint dates, and more concern about lead-tainted water. Similarly, owner occupied households had a negative correlation with the midpoint date. Since owner occupiers may tend to have more knowledge about their home infrastructure and may expect to remain in a location for longer than renters – perhaps even over generations – they could have a greater interest in ensuring low levels of water-based lead. The positive correlation of homes built between 1940-1959 may be due to a geographic anomaly since zip codes with the highest proportion of these homes are all in Eastern Queens. This region has very high median incomes which may ultimately explain the later midpoint dates.

This analysis indicates that more education and outreach to lower-income families by the New York City Department of Environmental Protection could be an effective means of addressing residents' concerns about future health risks. Additionally, it suggests an information disparity between renters and owner-occupiers that may be of interest to policy makers. 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 [149]. While some 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 midpoint 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.

In addition to the regression factors, there is a significant positive correlation between change slope and midpoint date with a p-value of 4×10^{-4}. The positive correlation between midpoint date and change slope is evident from a visual inspection of Figures 2.14 and 2.15. This relation indicates that in zip codes that changed later, their changes were relatively quicker perhaps due to the prevalence of news coverage at that later time.

2.4.4 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 [147]. 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 GPCS with two latent functions, spectral mixture kernels, and $w(x)$ defined by RKS features. We do not provide prior information about the 1963 vaccination date. Results for three states are shown in Figure 2.16 along with the predicted change surface for each state. The red line marks the vaccine year of 1963, while the magenta line marks where $\sigma(w(x_{\text{state}})) = 0.5$.

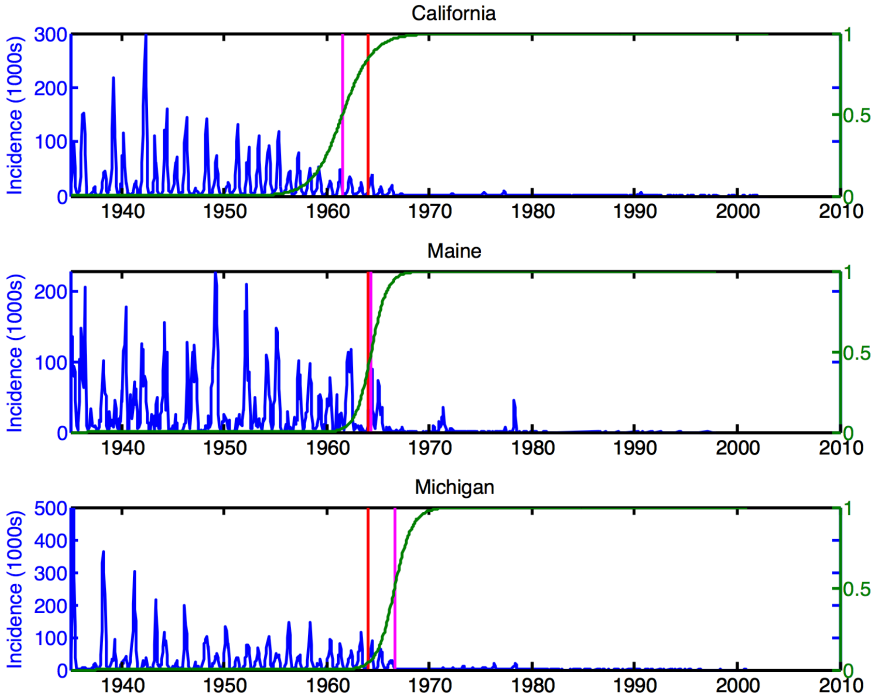


Fig. 2.16 Measles incidence levels from three states, 1935 to 2003. The green line plots $\sigma(w(x_{\text{state}}))$, the vertical red line indicates the vaccine in 1963, and the magenta line indicates $\sigma(w(x_{\text{state}})) = 0.5$.

GPCS 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. This finding again demonstrates the effectiveness of GPCS to detect changes in real world data while providing important insight into the change's dynamics that are not ascertainable through existing models. In Figure 2.17 we depict the midpoint, $\sigma(w(x_{\text{state}})) = 0.5$, 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 a later

midpoint in blue. We discover that there is an approximately 6 year range of midpoints between states, with California being the earliest and North Dakota being the latest.

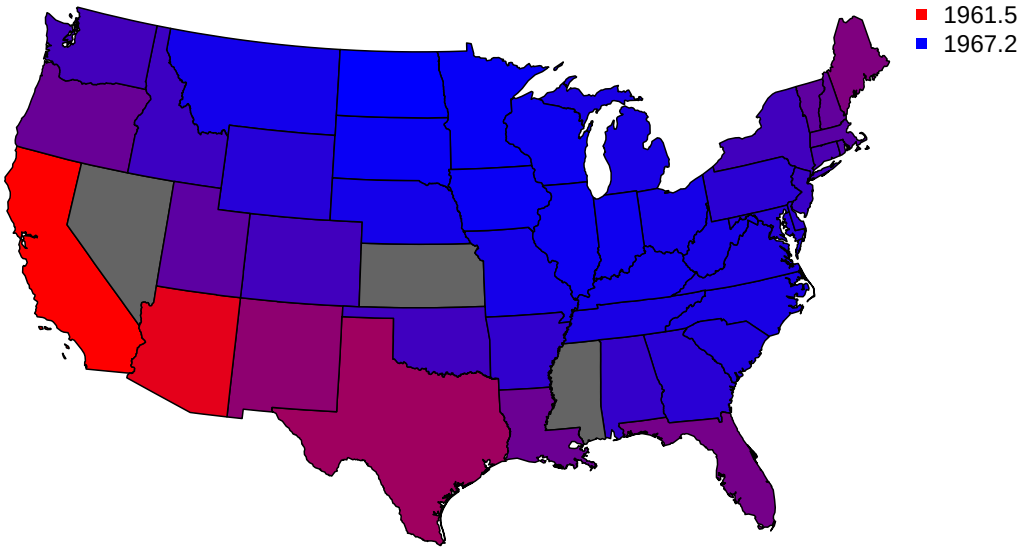


Fig. 2.17 U.S. states colored by the date where $\sigma(w(x_{\text{state}})) = 0.5$. Red indicates earlier dates, with California being the earliest. Blue indicates later dates, with North Dakota being the latest. Grayed out states were missing in the dataset.

In Figure 2.18 we depict the change surface slope from $\sigma(w(x_{\text{state}})) = 0.25$ to $\sigma(w(x_{\text{state}})) = 0.75$ 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 having the flattest slope and Maine the steepest.

Regression analysis: 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 midpoint shift date between the two regimes, $\sigma(w(x_{\text{state}})) = 0.5$. Since the change surface shifts primarily during the 1960s and the measles vaccine is introduced in 1963, we consider historical census data only from 1960-1962 [29]. Factors included annual birth rate, death rates of different age segments, and 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 has shown that birth rates and the size of a young non-immune population is important for

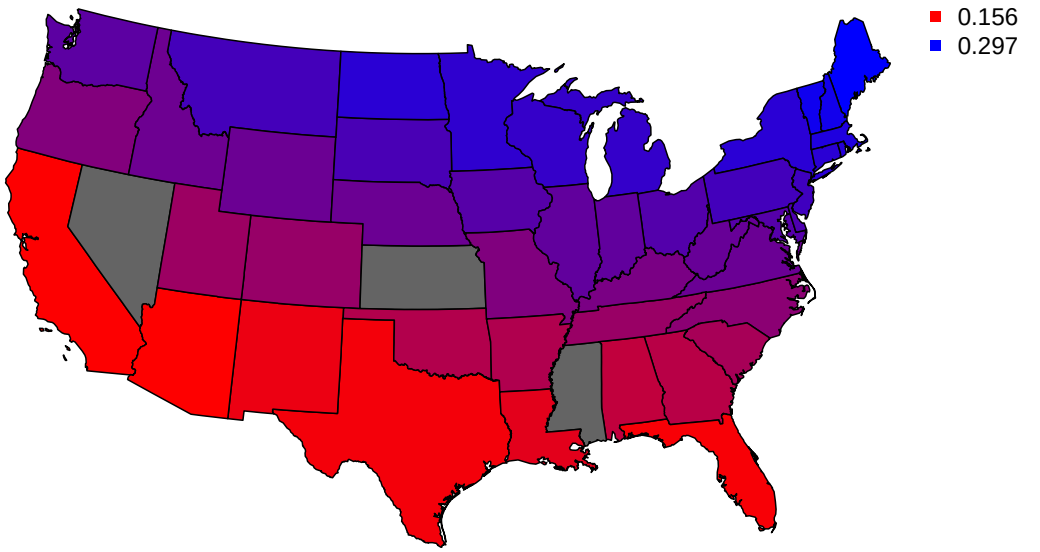


Fig. 2.18 U.S. states colored by the slope of $\sigma(w(x_{\text{state}}))$ from 0.25 to 0.75. Red indicates flatter slopes, with Arizona being the lowest. Blue indicates steeper slopes, with Maine being the highest. Grayed out states were missing in the dataset.

understanding the pre-vaccination dynamics of measles [46]. Indeed, before the measles vaccine 5-9 year olds comprised 50% of disease incidence [39]. We also consider median household income and household income inequality for each state. Finally, we also consider the average annual temperature in each state.

The results of a linear regression over all factors can be seen in Table 2.5. Four variables were statistically significant at a p-value < 0.05: the Gini coefficient of annual family income per state, average annual temperature, death rate of people aged 10+, and proportion of population aged 0-9. The Gini coefficient had a relatively large, positive correlation suggesting 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 midpoint 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 midpoint dates sometime in 1966. Another interesting example is the District of Columbia, which had the highest Gini coefficient. Although Washington D.C. is an urban center, it had also been an area of poverty and substandard local government, which may have contributed to its late change. Warmer temperatures are correlated with early midpoint dates perhaps

Table 2.5 Results from a linear regression to the measles incidence midpoint date, $\sigma(w(x_{\text{state}})) = 0.5$. Variables are listed on the left while their coefficients, with standard errors in parentheses, are listed on the right. Asterisks indicate statistically significant variables.

	Dependent variable:
	Midpoint date
Log death rate aged 0-4	−1.614
	(2.186)
Log death rate aged 5-9	5.023
	(2.640)
Log death rate aged 10+	7.651**
	(2.632)
Log birth rate	−10.932
	(5.472)
Gini of family income	48.503**
	(17.461)
Log median household income	4.997
	(2.620)
Log population	0.117
	(0.228)
Proportion of population aged 0-9	84.757*
	(32.784)
Average temperature (°F)	−0.093*
	(0.035)
Constant	1,980.509**
	(24.237)
Observations	46
R^2	0.396
Adjusted R^2	0.245
Note:	*p<0.05; **p<0.01

due to biological mechanisms underlying the contagion of measles. Additionally, measles is spread through human contact which may also be affected by weather patterns. Death rates of people aged 10+ and relatively larger populations of children aged 0-9 were associated with later midpoint dates. Both of these factors indicate higher density of young children who may never have been affected by measles. This in turn may have increased the prevalence of the virus and delayed the midpoint date.

In addition to the regression factors, there is a significant positive correlation between change slope and midpoint date with a p-value $< 2.2 \times 10^{-16}$, suggesting 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 midpoint 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. Indeed, one challenge in analyzing measles at a state-level aggregation is that measles disease dynamics may vary between cities even within states [40]. Nevertheless, the results 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. Additionally, care should be taken when accounting for the effects of weather patterns and population dynamics.

Chapter 3

Counterfactual Prediction with Change Surfaces

3.1 Introduction

As we observed while motivating the need for change surface models, interrogating the richness of complex changes helps us to understand those changes and describe what happened in the data. Yet these descriptions are fundamentally about past data. Policy makers often want to know, not only what happened in the past, but also what *could* have happened given different a different set of circumstance or changes. Such information provides critical insight for planning future policies and interventions. These types of "what if" questions are essentially counterfactual questions[1].

In this chapter we derive counterfactual prediction techniques using change surfaces. This allows a new framework for conceptualizing counterfactuals within the context of change analysis. In particular, the counterfactuals are computed with respect to real-valued labels, $s(x)$. Such labels provide a natural mechanism to encode partial treatment or spillover effects. Additionally, we derive counterfactual prediction for GPCS. Given the Bayesian nature of GPs this provides us with both posterior mean and covariance estimates for each point in the input domain. Finally, we demonstrate GPCS counterfactual prediction on *out-of-class* numerical data and complex spatiotemporal data.

[1] Published as Herlands et al. [63]

3.2 Counterfactual prediction for change surfaces

By simultaneously characterizing the change surface, $s(x)$, and the underlying generative functions, $f(x)$, change surface models allow us to ask questions about how the data would have looked had there been only one latent function. In other words, change surface models allow us to consider counterfactual questions.

For example, in Section 3.4.2 we consider measles disease incidence in the United States in the twentieth century. The measles vaccine was introduced in 1963, radically changing the dynamics of disease incidence. Counterfactual studies such as van Panhuis et al. [147] attempt to estimate how many cases of measles there would have been in the absence of the vaccine. To be clear, since change surface models do not consider explicit indicators of an intervention, they do not directly estimate the counterfactual with respect to a particular treatment variable such as vaccination. Instead, they identify and characterize changes in the data generating process that may or may not correspond to a known intervention. The change surface counterfactuals estimate the y values for each functional regime in the absence of the change identified by the change surface model. In cases where the discovered change surface does correspond to a known intervention of interest, domain experts may interpret the change surface predictions as a counterfactual "what if" that intervention and any contemporaneous changes in the data generating process (note that we cannot disentangle these causal factors without explicit intervention labels) did not occur.

Counterfactuals are typically studied in econometrics. In observational studies econometricians try to measure the effect of a "treatment" over some domain. Econometric models often measure simple features of the intervention effect, such as the expected value of the treatment over the entire domain, also known as the *average treatment effect*. A nascent body of work considers machine learning approaches to provide counterfactual prediction in complex data [17, 23, 79, 57], as well as richer measures of the intervention effect [17, 71]. Recent work by [131] uses Gaussian processes for trajectory counterfactual prediction over time. However, these methods generally follow a common framework using the potential outcomes model, which assumes that each observation is observed with a discrete treatment [126, 67]. With discrete treatments a unit, x, is either intervened upon or not intervened upon — there are no partial interventions. For example, in a medical study a patient may be given a vaccination, or given a placebo. Such discretization is similar to a traditional changepoint model where $s(x) \in \{0, 1\}$ can only be in one of two states. Yet discrete states prove challenging in practical applications where units may be partially treated or affected through spillover. For example, there may be herd effects in vaccinations whereby a person's neighbor being vaccinated reduces the risk of infection to the person. Certain econometric models attempt to account for partial treatment such as treatment eligibility [1], where partial

treatments are induced by defining proportions of the population that could potentially be treated. Yet a model that directly enables and estimates continuous levels of treatment may be more natural in such cases.

Counterfactuals using change surfaces. Change surface models enable counterfactual prediction in potentially complex data through the expressive parameterization of the latent functions, $f_1(x), \ldots, f_r(x)$. Determining the individual function value $f_i(x)$ over the input domain is equivalent to determining the counterfactual of $f_i(x)$ in the absence of all other latent functions. We can compute counterfactual estimates for latent functions in either the regular change surface model or the change surface background model. In the latter case, if $r = 2$ recall that the model takes the form $y = f_0(x) + s_1(x)f_1(x) + \varepsilon$. Determining the counterfactual for $f_0(x)$ provides an estimate for the data without the detected change, while the counterfactual for $f_1(x)$ estimates the effect of that change across the entire regime.

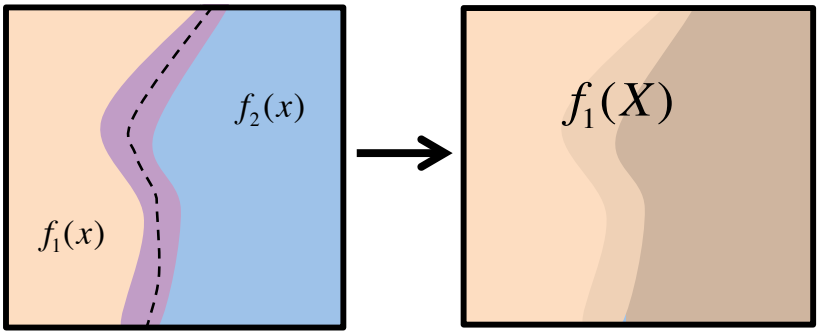


Fig. 3.1 Two-dimensional depiction of change surface counterfactual prediction. The left panel illustrates the change surface of Figure 2.1. The right image depicts the counterfactual of f_1 over the entire domain, X, representing what the observed data could look like in the *absence* of an intervention. The darker shading of the picture depicts larger posterior uncertainty.

For example, Figure 3.1 depicts the counterfactual of f_1 from Figure 2.1, where f_1 is predicted over the entire regime, X. The darker shading of the picture depicts larger posterior uncertainty. As we move toward the right portion of the plot, away from data regions where f_1 was active, we have greater uncertainty in our counterfactual predictions.

Computing counterfactuals for each $f_i(x)$ provides insight into the effect of a change on the various regimes. When combined with domain expertise, these models may also be useful for estimating the treatment effect of specific variables. Additionally, given a Bayesian formulation of the change surface, such as that proposed in Section 2.3, we can compute the full posterior distribution over the counterfactual prediction rather than just a point estimate.

Finally, since change surfaces model all data points as a combination of latent functions, we do not assume that observed data comes from a particular treatment or control. Rather we learn the contribution of each functional regime to each data point.

Some simple changepoint models could, in theory, provide the ability for counterfactual prediction between regimes. But since changepoint models consider each regime either completely or nearly independently of other regimes, there is no information shared between regimes. This lack of information sharing across regimes makes accurate counterfactual prediction challenging without strong assumptions about the data generating process. Indeed, to our knowledge there is no previous literature using changepoint models for counterfactual prediction.

Assumptions in change surface counterfactuals: Change surface models identify changing data dynamics without explicitly considering intervention labels. Instead, counterfactuals of the functional regimes are computed with respect to the change surface labels, $s(x)$. Thus these counterfactuals estimate the value of functional regimes in the absence of those changes but do not necessarily represent counterfactual estimates of any particular variable. The interpretation of these counterfactuals as estimates for each functional regime in the absence of a specific known intervention requires identification of the correct change surface, i.e.:

- The intervention induces a change in the data generation process that cannot be modeled with a single latent functional regime.
- The magnitude of the change is large enough to be detected.
- The change surface model is sufficiently flexible to accurately characterize this change.
- The change surface model does not overfit the data to erroneously identify a change.

Moreover, the resulting counterfactual estimates do not rule out the possibility that other changes in the data generating process occur contemporaneously with the intervention of interest. As such, these counterfactuals are most naturally interpreted as estimating what the data would look like in the absence of the intervention and any other contemporaneous changes. Disentangling these multiple potential causal factors would require additional data about both the intervention and other potential causes.

Change surface counterfactual predictions can provide immense value in practical settings. Although in some datasets explicit intervention labels are available, many observational datasets do not have such labels. Learning a change surface effectively provides a real-valued label that can be used to predict counterfactuals. Even when the approximate boundaries of an intervention are known, change surface modeling can still provide an important advantage

since the intervention labels may not capture the true complexity of the data. For example, knowing the date that the measles vaccine was introduced does not account for regional variation in vaccine distribution and uptake (see Section 3.4.2). Both observational studies and randomized control trials suffer from partial treatment or spillover, where an intervention on one agent or region secondarily affects a non-intervened agent or region. For example, increasing policing in one area of the city may displace crime from the intervened region to other areas of the city [148]. This effect violates the Stable Unit Treatment Value Assumption, which is the basis for many estimation techniques in economics [125]. By using the assumed boundaries of an intervention as a prior over $s(x)$, a change surface model can discover if, and where, spillover occurs. This spillover will be captured as a non-discrete change and can aid both in interpretability of the results and counterfactual prediction. In all these cases change surface counterfactuals may lead to more believable counterfactual predictions by using a real valued change surface to directly model spillover and interventions.

3.3 GPCS Counterfactual Prediction

We consider counterfactuals when using two latent functions in a GPCS, $f_1(x)$ and $f_2(x)$. This two-function setup addresses a typical setting for counterfactual prediction when considering two alternatives. The derivations below can be extended to multiple functional regimes. As discussed above, we note that change surface counterfactuals are only valid with respect to the regimes of the data as identified by GPCS. Subsequent analysis and domain expertise are necessary to make any further claims about the relationship between an identified change surface and some latent intervention.

In counterfactual prediction we wish to infer the value of $f_1(x)$ and $f_2(x)$ in the absence of the other function. Therefore we condition on the observations, (x,y), and GPCS model parameters in order to compute the conditional distribution $p\big([f_1(x),f_2(x)]|y\big)$ from the multivariate Gaussian joint distribution $p\big([f_1(x),f_2(x)],y\big)$. For notational convenience we omit explicit reference to the model parameters in the subsequent derivations but note that all distributions are conditional on these parameters.

To recall, for two latent functions, $f_1(x)$ and $f_2(x)$, GPCS specifies

$$y(x) = \sigma_1 f_1(x) + \sigma_2 f_2(x) + \varepsilon \tag{3.1}$$

$$\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2) \tag{3.2}$$

$$f_1(x) \sim \mathcal{GP}(0, K_1) \tag{3.3}$$

$$f_2(x) \sim \mathcal{GP}(0, K_2) \tag{3.4}$$

where for notational simplicity we let $K_1 = k_1(x,x')$, $K_2 = k_2(x,x')$, $\sigma_1 = \sigma_1(w(x))$, and $\sigma_2 = \sigma_2(w(x))$.

We consider the most general case when we want to predict counterfactuals for both $f_1(x)$ and $f_2(x)$ over the domain X. No restrictions are placed over X. It can include the entire original domain, parts of the original domain, or different inputs entirely. We concatenate $f_1(X)$ and $f_2(X)$ together,

$$u = [f_1(X), f_2(X)]. \tag{3.5}$$

Since in Section 2.3.1 we assumed that $f_1(x)$ and $f_2(x)$ have independent Gaussian process priors, we know that,

$$u \sim \mathcal{N}\left(0, \begin{bmatrix} K_1 & 0 \\ 0 & K_2 \end{bmatrix}\right) \tag{3.6}$$

Considering the observed data, y, we know that u and y are jointly Gaussian,

$$\begin{bmatrix} u \\ y \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \Sigma_{u,u} & \Sigma_{u,y} \\ \Sigma_{u,y}^T & \Sigma_{y,y} \end{bmatrix}\right) \tag{3.7}$$

and using multivariate Gaussian identities, we find that u has the conditional Gaussian distribution

$$u|y \sim \mathcal{N}\left(\Sigma_{u,y}\Sigma_{y,y}^{-1}y, \Sigma_{u,u} - \Sigma_{u,y}\Sigma_{y,y}^{-1}\Sigma_{u,y}^T\right) \tag{3.8}$$

Thus in order to derive counterfactuals for both $f_1(X)$ and $f_2(X)$ we only need to compute $\Sigma_{u,y}, \Sigma_{y,y}$, and $\Sigma_{u,u}$. Note that with respect to $\Sigma_{u,u}$ we have already derived the covariance structure for u in Equation (3.6).

Computation for $\Sigma_{u,y}$ In order to compute $\Sigma_{u,y}$, we expand the multiplication noting that y is defined to be a two-function GPCS,

$$\Sigma_{u,y} = E[uy^T] \tag{3.9}$$

$$= \mathbb{E}[\begin{bmatrix} f_1(x_1) \\ \dots \\ f_1(x_n) \\ f_2(x_1) \\ \dots \\ f_2(x_n) \end{bmatrix} \begin{bmatrix} \sigma_1(x_1)f_1(x_1) + \sigma_2(x_1)f_2(x_1) + \varepsilon \\ \dots \\ \sigma_1(x_n)f_1(x_n) + \sigma_2(x_n)f_2(x_n) + \varepsilon \end{bmatrix}^T] \tag{3.10}$$

Multiplying these elements is assisted by the following identities. Recall that kernels K_1 and K_2 define the covariance among function values in f and g respectively,

$$\mathbb{E}[f_1(x_i)f_1(x_j)] = k_1(i,j) \tag{3.11}$$

$$\mathbb{E}[f_2(x_i)f_2(x_j)] = k_2(i,j) \tag{3.12}$$

Additionally, since $f_1(x)$ and $f_2(x)$ have independent Gaussian process priors, $\mathbb{E}[f_1(x_i)f_2(x_j)] = 0$. Furthermore, because ε is distributed with mean zero, $\mathbb{E}[\varepsilon_i] = 0$. Finally, since $\sigma_1(x)$ and $\sigma_2(x)$ are constant (conditional on hyperparameters) $\mathbb{E}[\sigma_1(x_i)] = \sigma_1(x_i)$ and $\mathbb{E}[\sigma_2(x_i)] = \sigma_2(x_i)$. Thus we can conclude that

$$\Sigma_{u,y} = \begin{bmatrix} \sigma_1(x_1)k_1(1,1) & \sigma_1(x_2)k_1(1,2) & \dots & \sigma_1(x_n)k_1(1,n) \\ \sigma_1(x_1)k_1(2,1) & \sigma_1(x_2)k_1(2,2) & \dots & \sigma_1(x_n)k_1(2,n) \\ \dots & & & \\ \sigma_1(x_1)k_1(n,1) & \sigma_1(x_2)k_1(n,2) & \dots & \sigma_1(x_n)k_1(n,n) \\ \sigma_2(x_1)k_2(1,1) & \sigma_2(x_2)k_2(1,2) & \dots & \sigma_2(x_n)k_2(1,n) \\ \sigma_2(x_1)k_2(2,1) & \sigma_2(x_2)k_2(2,2) & \dots & \sigma_2(x_n)k_2(2,n) \\ \dots & & & \\ \sigma_2(x_1)k_2(n,1) & \sigma_2(x_2)k_2(n,2) & \dots & \sigma_2(x_n)k_2(n,n) \end{bmatrix} \tag{3.13}$$

$$= \begin{bmatrix} K_1 \odot \mathbb{1}\sigma_1^T \\ K_2 \odot \mathbb{1}\sigma_2^T \end{bmatrix} \tag{3.14}$$

where $\odot$ is elementwise multiplication.

Computation for $\Sigma_{y,y}$ The computation for $\Sigma_{y,y}$ is very similar to that of $\Sigma_{u,y}$ so we omit its expansion for the sake of brevity. The slight difference is that we must consider $\mathbb{E}[\varepsilon_i \varepsilon_i]$ which equals σ_ε^2.

Thus,

$$\Sigma_{y,y} = E[yy^T] \tag{3.15}$$
$$= K_1 \odot [\sigma_1 \sigma_1^T] + K_2 \odot [\sigma_2 \sigma_2^T] + I_n \sigma_\varepsilon^2 \tag{3.16}$$

GPCS background model counterfactuals

The counterfactual derivations above directly apply to the GPCS background model with $r = 2$, where $y(x) = f_0(x) + \sigma_1(w(x)) f_1(x)$. Recall that as we discussed in Section 2.2.1, this is a special case of the GPCS background model where $f_1(x)$ is an additive change function. In this case, the counterfactual for $f_0(x)$ estimates what would have occurred in the absence of the identified change. The counterfactual for $f_1(x)$ models how the change would have affected the entire domain.

If we let $u = [f_0(X), f_1(X)]$ we can derive counterfactuals for the GPCS background model by setting $\sigma_0 = 1$ in the equations for $\Sigma_{u,u}$, $\Sigma_{u,y}$, and $\Sigma_{y,y}$ above. Explicitly,

$$\Sigma_{u,u} = \begin{bmatrix} K_0 & 0 \\ 0 & K_1 \end{bmatrix} \tag{3.17}$$
$$\Sigma_{u,y} = \begin{bmatrix} K_0 \\ K_1 \odot \mathbb{1}\sigma_1^T \end{bmatrix} \tag{3.18}$$
$$\Sigma_{y,y} = K_0 + K_1 \odot [\sigma_1 \sigma_1^T] + I_n \sigma_\varepsilon^2 \tag{3.19}$$

3.4 Experiments

Using two-dimensional numerical data in Section 3.4.1 we compute highly accurate counterfactual predictions for both GPCS and GPCS background models and discuss how the posterior distribution varies over the prediction domain as a function of the change surface. Additionally, using the US measles data from 2.4.4, we estimate the counterfactual of measles incidence without vaccination by filtering out the detected change function and examining the inferred latent background function.

3.4.1 Numerical Experiments

We use GPCS to compute counterfactual predictions on the numerical data. In the previous experiments from Section 2.4 we used the data, (x, y), to fit the parameters of GPCS, θ. Now we condition on (x, y, θ) to infer the individual latent functions $f_1(x)$ and $f_2(x)$ over the entire domain, x. By employing the marginalization procedure described in Section 3.3 we derive posterior distributions for both $f_1(x)$ and $f_2(x)$. Since we have synthetic data we can then compare the counterfactual predictions to the true latent function values. Specifically, we use (x, y, θ) from Figure 2.7b to infer the posterior counterfactual mean and variance for both $f_1(x)$ and $f_2(x)$ and show the results in Figure 3.2. Note how the posterior mean predictions

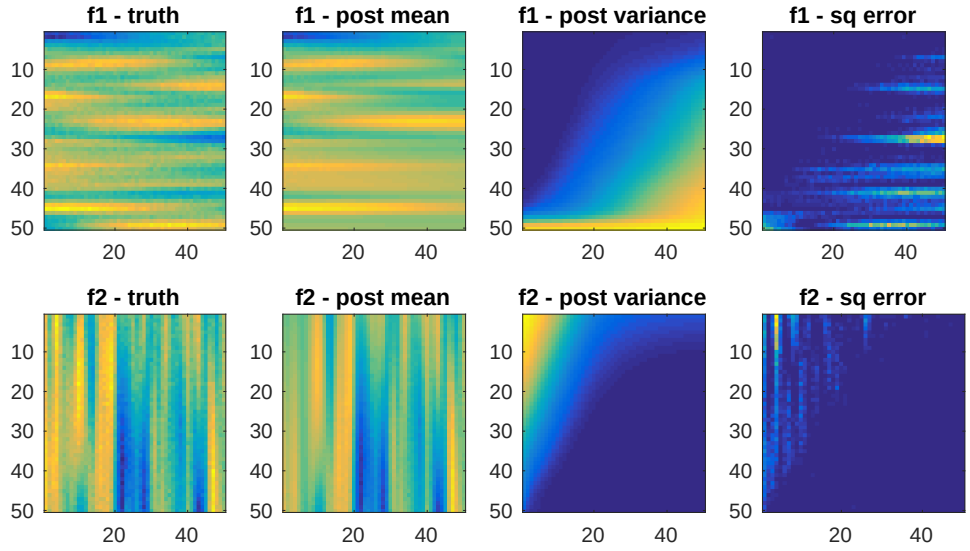


Fig. 3.2 Posterior counterfactual predictions using hyperparameters derived from GPCS model. We plot the true latent function as well as the posterior mean and variance estimates for each function. Additionally, we plot the squared error between the true and posterior mean values.

of $f_1(x)$ and $f_2(x)$ are quite similar to the true values. Moreover, the posterior uncertainty estimates are very reasonable. For both $f_1(x)$ and $f_2(x)$ the posterior variance varies over the two-dimensional domain, x, as a function of the change surface. Where $s_1(x) \approx 1$ the posterior variance of $f_1(x) \approx 0$ while the posterior variance of $f_2(x)$ is large. In areas where $s_2(x) \approx 1$ the posterior variance of $f_1(x)$ is large, while the posterior variance of $f_2(x) \approx 0$. The uncertainty is also evident in the squared error, $\frac{1}{n}\sum(f_i(x) - f_i\hat{(x)})$, where, as expected, each function has larger error in areas of high posterior variance.

GPCS background model

We use the GPCS background model to compute counterfactual predictions on the data from Figure 2.10b. Conditioning on (x, y, θ) we employ the marginalization procedure described in Section 3.3 to infer posterior distributions for the background function, $f_0(x)$, and the change function, $f_1(x)$, over the entire domain, x. The results are shown in Figure 3.3. Note

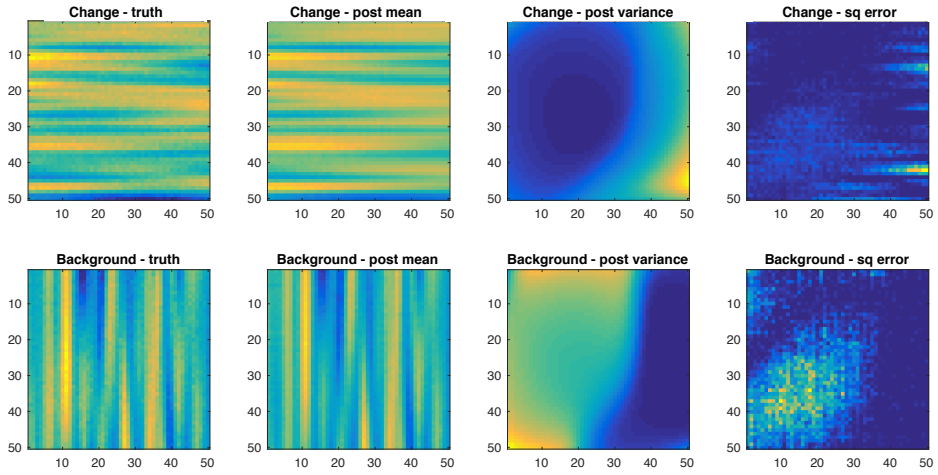


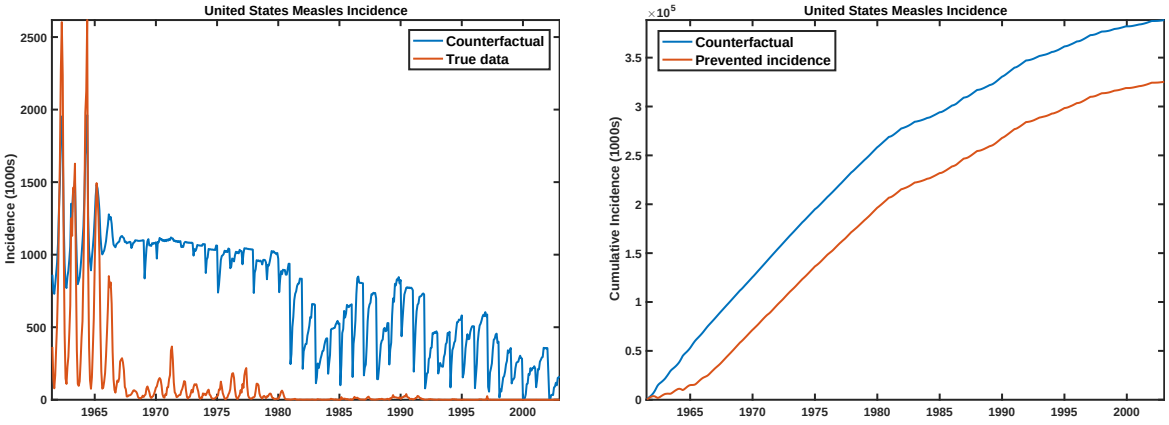
Fig. 3.3 Posterior counterfactual predictions using hyperparameters derived from GPCS background model. We plot the true latent function as well as the posterior mean and variance estimates for each function. Additionally, we plot the squared error between the true and posterior mean values.

how the posterior mean predictions of both the background and change functions are quite similar to the true values. As in the case of GPCS, the posterior variance for each function varies over the two-dimensional domain, x, as a function of the change surface, $\sigma(w_{\text{poly}}(x))$.

3.4.2 United States Measles Data Counterfactuals

Using the counterfactual GPCS framework and data from Section 2.4.4, we inferred the incidence of measles in the absence of the change surface identified by GPCS. We used the latent function that is dominant in the data before the measles vaccine to compute posterior estimates for measles incidence between the earliest detected midpoint date in 1961 and the end of the data in 2003. This estimation is inspired by the counterfactual estimation described in van Panhuis et al. [147]. We argue that GPCS provides more believable counterfactual estimates than simple interpolations or regressions because GPCS is a more expressive model for measles dynamics and explicitly considers data variation both before and after the start of the measles vaccine program. Figure 3.4 depicts the aggregated counterfactual posterior

mean estimates over the entire United States. The left plot shows true and counterfactual monthly incidence, while the right plot depicts the cumulative counterfactual incidence. Under the assumption that the change surface reflects the causal effect of the vaccine program intervention, we also estimate how many cases were "prevented" through the vaccination program. Since disease dynamics may have many causal factors, we cannot disentangle the introduction of the measles vaccine from any contemporaneous societal or policy changes that may have impacted measles incidence. Thus these findings are a starting point for more extensive epidemiological research. Additionally, while we plot the posterior mean estimates, note that our confidence in these estimates diminishes as we consider counterfactual estimates far from the change region.



(a) (b)

Fig. 3.4 Counterfactual posterior mean estimates for measles incidence. Plot (a) depicts the aggregated counterfactual posterior mean estimates over the entire United States. Plot (b) depicts the cumulative counterfactual incidence over the entire United States as well as estimating how many cases were "prevented" through the vaccination program under the assumption that the change surface corresponds to the vaccine intervention.

Chapter 4

Anomalous Pattern Detection in Non-iid Data

4.1 Introduction

Anomalous pattern detection is the task of identifying subsets of data points that systematically differ from the underlying model. Identifying anomalous patterns in real-world data is critical for understanding how people and systems deviate from expected behavior. In the spatiotemporal domain, timely identification of such patterns can allow for effective interventions. For example, detecting anomalous increases in opioid deaths can enable health care workers to effectively target overdose prevention programs. Similarly, patterns of increased 311 calls can help cities to better target services and allocate resources[1].

To detect these anomalous patterns, we will address three key challenges. First, real-world data is extremely complex with non-trivial correlations across space, time, and other features. Treating data points as iid ignores important covariance structure and will substantially overestimate the anomalousness of detected patterns. Second, an event of interest often affects multiple nearby points. Simply considering how anomalous is each individual point loses power to detect subtle anomalies. Third, anomalous patterns are often irregularly shaped or discontiguous due to latent demographic or geographic features. Searching for these complex patterns is important for precision and detection power, yet exhaustive methods are computationally intractable and may result in overfitting.

A sensible approach to this problem is model-based anomaly detection, where a distribution is fit to model "regular" data. Points with a low likelihood under this distribution are identified as anomalous [32, 66]. To address the complex correlations in real-world systems,

[1] Published as Herlands et al. [61]

Gaussian processes (GPs) provide a natural means of learning covariance structure from data. However, GP anomaly detection has been typically used to classify *individual* points as outliers [138, 84, 140]. Such approaches have difficulty when confronted with subtle anomalies, where each individual data point may seem to conform to the underlying distribution, yet when taken as a group, they form a collectively anomalous pattern. Thus anomalous pattern detection is a conceptually and statistically different problem than anomaly or outlier detection.

A few recent GP models consider anomalous intervals [118] and sophisticated change points [128, 64] to detect intervals of anomalous points. However, these methods (the first two of which are applied exclusively to one-dimensional data) are limited to contiguous intervals in the input domain and cannot model the irregularly shaped anomalies we expect in complex data. Cheng et al. [34] recently developed an anomalous pattern detection technique for spatiotemporal data. However, this approach requires a corpus of anomaly-free training data, can only detect contiguous anomalous patterns, and is specific to video data.

In the statistics literature, spatial and subset scanning methods are commonly used to identify collectively anomalous subsets of data [85, 102]. By combining information across a subset of data elements, they generate a strong signal of anomalous behavior. These approaches compute a log-likelihood ratio (LLR) of subsets being drawn from a null or anomalous distribution. The LLR is a powerful statistic that measures how much evidence exists in the data to conclude if the subset exhibits abnormal behavior [85, 105]. A core challenge of subset scanning is searching through the $O(2^n)$ possible subsets of n data elements [104, 7, 45, 159]. Neill [102] shows that certain LLR statistics satisfying a linear-time subset scanning (LTSS) property can be optimized in $O(n \log n)$ by ordering points according to a particular "priority function" and evaluating only n of the 2^n subsets. However, LTSS assumes that we can compute the contributions of individual points to the LLR. This is possible only when assuming that data is uncorrelated under the null, as in the case of an independent Gaussian scan statistic [101]. Yet when applied to non-iid data this independence assumption would result in substantial false positive rates since correlated fluctuations will be mistaken for anomalous movements.

4.1.1 Contributions

In this chapter we introduce novel techniques for identifying anomalous patterns in non-iid data. Our methods are powerful and interpretable. By combining naturally interpretable GPs with localized anomalous patterns we can describe the "regular" data dynamics as well as quantify and corroborate anomalous regions with domain experts. Our main contributions are:

1. Combining GP modeling with subset scanning for powerful and interpretable detection of anomalous patterns in highly correlated data.

2. Proposing a new likelihood ratio statistic and subset scan technique for correlated data that do not assume conditional independence.

3. Performing hold-out GP inference while computing our new likelihood ratio statistic conditioned on GP hyperparameters, to avoid corrupting the null model with anomalies.

4. Developing two novel, principled approaches to the NP-hard problem of searching for the most anomalous subset, through a new iterative method and an application of the Generalized Rayleigh Quotient respectively.

5. We demonstrate our methods on numeric simulations, opioid-related deaths, 311 calls for service data, and multiple streams of sewer flooding reports and tree damage reports, illustrating interpretable and policy-relevant results.

The chapter proceeds as follows: Section 4.2 introduces a novel log-likelihood ratio statistic for non-iid data. Section 4.3 details the Gaussian Process Neighborhood Scan (GPNS) and the Gaussian Process Subset Scan (GPSS). Experimental results on numerical and real data are presented in Section 4.4.

4.2 LLR statistic for non-iid data

Consider data, (x,y), where $x = \{x_1,\ldots,x_n\}, x_i \in \mathbb{R}^D$, are inputs or covariates, and $y = \{y_1,\ldots,y_n\}$, $y_i \in \mathbb{R}$ are outputs or response variables indexed by x. We are interested in anomalous patterns that systematically differ from the underlying data distribution. We frame this search as an LLR comparison between a null model of "regular" behavior and an alternative model of "anomalous" behavior. A single latent GP defines both models. Subsets of data with the highest LLR scores are identified as the most anomalous and randomization testing identifies a threshold for statistical significance.

Using a GP as the foundational modeling technique enables us to learn complex covariance structure and seamlessly extend to high dimensions as well as missing data. GPs are also naturally interpretable, which can provide insight about the "regular" data dynamics.

Consider a given subset of data points defined by the binary weighting vector w, where $w_i = 1$ if (x_i, y_i) is included in the subset and $w_i = 0$ if excluded. Our null model, H_0, assumes that all points (regardless of w_i) are drawn from a function with a GP prior: $y = f(x) + \varepsilon$, where $f(x) \sim GP(\theta_0)$ and $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 I)$. Our alternative model, $H_1(w)$, assumes that

$y_i = f(x_i) + \varepsilon$ for $w_i = 0$, and $y_i = g(f(x_i), \theta_1) + \varepsilon$ for $w_i = 1$, where $g(\cdot)$ is any function of the latent GP.

Here we focus on the case of a mean shift, $g(f(x), \theta_1) = f(x) + \beta$, $\beta \in \mathbb{R}^1$. The covariance structure remains the same in the null and alternative models. This allows us to efficiently compute the posterior mean vector μ and covariance matrix Σ through GP inference, where $y \sim \mathcal{N}(\mu, \Sigma)$ under H_0, and $y \sim \mathcal{N}(\mu + \beta w, \Sigma)$ under $H_1(w)$. For posterior μ and Σ we condition on all data outside the subset of points represented by w, ensuring that null model estimates are not corrupted by anomalous observations. However, since anomalies are assumed to be rare, their influence on parameter estimation is minimal. Therefore we use all (x, y) for GP learning of the parameters of the null model θ_0.

We concentrate on mean changes since many real world cases concern anomalous levels of a quantity. Increases in localized drug overdoses, crime, and calls for city service are all mean shifts of great importance. Methods for identifying arbitrary changes in distribution – while able to detect other sorts of patterns – have reduced power to detect such mean shifts, due to more diffuse inductive biases. Persistent changes in covariance structure are typically considered changepoints and require substantial data in both regimes as opposed to the localized anomalous patterns we detect.

To measure how anomalous is a subset defined by w, we compute the generalized log-likelihood ratio, $LLR(w) = \max_\beta LLR(w \,|\, \beta)$, where:

$$LLR(w \,|\, \beta) = \log \frac{\texttt{MNPDF}(y - \beta w \,|\, \mu, \Sigma)}{\texttt{MNPDF}(y \,|\, \mu, \Sigma)} \tag{4.1}$$

Here `MNPDF` is the multivariate normal probability density function. The most anomalous subset, w^*, is

$$\begin{aligned} w^* &= \operatorname*{argmax}_w LLR(w) \\ &= \operatorname*{argmax}_w \max_\beta -\frac{\beta^2}{2} w^T E w + \beta w^T E (y - \mu) \end{aligned} \tag{4.2}$$

where $E = \Sigma^{-1}$ for notational brevity. Conditional on w, we can determine the optimal mean shift, β^* through maximum likelihood estimation as shown below.

$$\begin{aligned}\beta^* &= \max_{\beta}\Big((2\pi)^{-\frac{k}{2}}|\Sigma|^{-\frac{1}{2}}\exp(-\frac{1}{2}(y-w\beta-\mu)^T \\ &\qquad E(y-w\beta-\mu))\Big) \\ &= \max_{\beta} -\frac{1}{2}(y-w\beta-\mu)^T E(y-w\beta-\mu) \\ &= \max_{\beta}(y-\mu)^T Ew\beta - \frac{1}{2}(w\beta)^T E(w\beta)\end{aligned} \tag{4.3}$$

We take the derivative with respect to β and set it to zero

$$\begin{aligned}\frac{\delta LLR(w)}{\delta\beta} &= (y-\mu)^T Ew - (w\beta^*)^T E(w) = 0 \\ &\Rightarrow (w\beta^*)^T E(w) = (y-\mu)^T Ew \\ &\Rightarrow \beta^* = \frac{w^T E(y-\mu)}{w^T Ew}\end{aligned} \tag{4.4}$$

Although β^* can be calculated in closed form, nevertheless, maximizing $LLR(w)$ is an NP-complete Integer Quadratic Program [41], so an optimal solution requires exponential-time computation. Note that the LTSS condition for a log linear-time subset search described in Neill [102] does not apply, since it requires independent data with a diagonal covariance matrix.

4.2.1 Randomization testing

Given a method for finding anomalous subsets, the following randomization testing procedure determines an α-level significance threshold for $LLR(w)$ conditional on the parameters of the null model:

1. Repeatedly draw $y^{(r)} \sim GP(\theta_0)$, at the same covariates, x, as the real data for $r = 1...R$.
2. Scan over $(x, y^{(r)})$ with the chosen subset searching method. For each randomization r save the most anomalous LLR value, $LLR(w^{*,(r)})$.
3. Determine an α-level threshold for significance based on the $(1-\alpha)$ quantile of the R maximum LLR values, above which any $LLR(w)$ from the original scan is considered statistically significant.

4.3 Efficient subset scanning

Having defined the LLR scan statistic to evaluate how anomalous is a given subset, we must now decide over which subsets to scan. Unconstrained optimization over $O(2^n)$ subsets is computationally infeasible for an exhaustive search. Additionally, an unconstrained search may return an unrelated set of points, reducing interpretability and increasing the potential for overfitting. Anomalous events in human data, such as drug usage and requests for government services, often affect multiple nearby points. Thus we assume that anomalous points are near one another. For example, in spatiotemporal data we assume that anomalous points are clustered in space and time. Following Neill [102], we define the local "k-neighborhood" of each data point, consisting of that point and its $k-1$ nearest neighbors, for some k. We propose two approaches for using these neighborhoods to identify anomalous patterns: Gaussian Process Neighborhood Scan (GPNS) and Gaussian Process Subset Scan (GPSS).

4.3.1 GP Neighborhood Scan (GPNS)

Given a maximum neighborhood size k_{max}, GPNS searches over the $O(nk_{max})$ local neighborhoods consisting of the k-neighborhood for each point where $k = \{1,2,\ldots,k_{max}\}$. Where neighborhoods are defined by Euclidean distance, such as in spatial data, the set of search regions are circular in shape. For each neighborhood, $(x^{(n)}, y^{(n)})$, we obtain posterior μ and Σ conditional on θ_0 and points $(x^{(-n)}, y^{(-n)})$. We then compute $LLR(w)$ for the neighborhood where $w = \vec{1}$, i.e., we evaluate the alternative hypothesis of the entire neighborhood being anomalous. GPNS pseudocode is presented in Algorithm 3.

Algorithm 3 GPNS

for $k = 1 : k_{max}$ **do**
  for (x_i, y_i), $i = 1 : n$ **do**
    Define k-neighborhood, $n^{(k,i)}$, and infer (μ, Σ)
    Set $w^{(k,i)} = \vec{1} \in \{0,1\}^k$
    Compute β^* given $w^{(k,i)}$
    Compute $LLR(w^{(k,i)})$
  end for
end for
Choose $n^* = \text{argmax}_{n^{(k,i)}} LLR_{n^{(k,i)}}$
Randomization testing for significance

4.3.2 GP Subset Scan (GPSS)

While GPNS simplifies the exponential search, it requires constraining assumptions about the shape of neighborhoods and is only able to discover contiguous, spherical anomalous patterns. While there are approaches to increase the variety of neighborhood shapes without substantially degrading computational efficiency [87, 104, 86], these methods still require strict specification of potential anomalies. Such foreknowledge is unrealistic in real-world applications where natural boundaries, demographics, and stochastic effects lead to irregularly-shaped patterns. In such cases GPNS has reduced detection and explanatory power.

To flexibly detect irregularly-shaped patterns, GPSS conducts an unconstrained search for the most anomalous subset within neighborhoods of fixed size k. Specifically, we identify the subset of points $(x^{(s)}, y^{(s)}) \subseteq (x^{(n)}, y^{(n)})$ that maximize the LLR within each neighborhood. This allows us to identify highly irregular and even non-contiguous anomalous patterns. By restricting the search within a local neighborhood, we ensure that the identified patterns are coherent and interpretable. GPSS requires evaluating $O(n)$ neighborhoods, as presented in Algorithm 4.

Algorithm 4 GPSS

1: Fix k at some size
2: **for** (x_i, y_i), $i = 1 : n$ **do**
3: Define k-neighborhood, $n^{(i)}$, and infer (μ,Σ)
4: Approximate the optimal subset, $s^{(i)} \subseteq n^{(i)}$
5: Set each $w_j^{(i)} = 1(j \in s^{(i)})$
6: Compute β^* given $w^{(i)}$
7: Compute $LLR(w^{(i)})$
8: **end for**
9: Choose $s^* = \text{argmax}_{s^{(i)}} LLR_{s^{(i)}}$
10: Randomization testing for significance

Unfortunately, this procedure requires finding $w \in \{0,1\}^k$ that maximizes the LLR of a subset within the neighborhood, argmax_w -$\frac{1}{2} w^T \beta E w \beta + w^T \beta E (y^{(n)}$-$\mu)$. This is still an Integer Quadratic Program, whose optimal solution is intractable even for moderately sized neighborhoods. Instead, below we formulate three approaches for finding approximate solutions.

β_{MAX} for conditionally optimal subset

Due to the full rank covariance matrix, we are unable to disentangle the individual contributions from each point to the LLR. However, if we condition on some subset of points, w, we are able to compute the conditional contribution of each point. First, note that conditional on w we can decompose w^* from Equation 4.2 into a sum over each of the m points in the neighborhood

$$\begin{aligned} & w^T \beta E(y^{(n)} - \mu) - \frac{1}{2} w^T \beta E w \beta \\ & = \sum_i w_i \Big[\beta \big(E(y^{(n)} - \mu) \big)_i - \frac{1}{2} \big(\sum_{j \neq i} w_j E_{j,i} + E_{i,i} \big) \beta^2 \Big] \end{aligned} \tag{4.5}$$

The contribution of point (x_i, y_i) to the LLR is the difference in LLR between $w_i = 0$ and $w_i = 1$. Due to the outer and inner sums, the change in the LLR is:

$$\beta \big(E(y^{(n)} - \mu) \big)_i - \frac{1}{2} \big(\sum_{j \neq i} 2 w_j E_{j,i} + E_{i,i} \big) \beta^2 \tag{4.6}$$

To maximize the LLR a point is only added to the subset if its contribution is positive. By setting Equation 4.6 to zero we can compute β_{MAX_i}, the maximum β value for which to include point (x_i, y_i).

$$\beta_{MAX_i} = \Big[2 \big(E(y^{(n)} \text{-} \mu) \big)_i \Big] / \Big[\sum_{j \neq i} 2 w_j E_{j,i} + E_{i,i} \Big] \tag{4.7}$$

As proved in Speakman et al. [139], we obtain the conditional optimal subset by using β_{MAX} as a priority function, ranking each data point by β_{MAX_i}, and iteratively compute the score function for subsets including each additional point. This yields a log linear search over data points. Such an approach identifies the most anomalous subset with a positive mean shift. To find the most anomalous subset with a negative mean shift we simply rank data points by $-\beta_{MAX_i}$

Since the derivation of β_{MAX_i} is conditional on a subset w, we obtain the *conditional* optimal subset. In order to approximate an optimal solution we use iteratively compute the conditional optimal subset beginning with a null subset, $w = \vec{0}$. This is an $O(\ell k \log(k))$ algorithm for some ℓ number of iterations, where k is the size of the neighborhood. Pseudocode is depicted in Algorithm 5.

For a diagonal Σ, β_{MAX} orders points according to $2(y_i^{(n)} - \mu_i)$, which is equivalent to the LTSS priority function for an independent Gaussian subset scan [139]. Thus β_{MAX} approach identifies the optimal subset in the independent case and is conditionally optimal in the dependent case.

Algorithm 5 Iterative β_{MAX_i} algorithm

1: Initialize $w = \vec{0}$
2: **for** $l = 1 : \ell$ **do**
3: Compute $\beta_{MAX_i} \forall i$ conditioned on the current value of w
4: Find highest scoring subset, $w^{(l)}$, using a linear search over sorted β_{MAX_i}
5: Compute $LLR(w^{(l)})$
6: Set $w = w^{(l)}$
7: **end for**
8: Choose $w^* = \text{argmax}_{w^{(l)}} LLR(w^{(l)})$

Although we focus on unconstrained subsets searching within neighborhoods, real world applications sometimes require a more constrained optimization. For example, in spatiotemporal phenomena it is often useful to consider anomalous patterns that are nearby in space and contiguous over time. We can enforce such constraints by predefining mutually exclusive blocks of points, $(x^{(B)}, y^{(B)}) \subseteq (x^{(n)}, y^{(n)})$ where points in a block must all either be included in, or excluded from, a subset.

When considering blocks of points we can compute the total contribution from all points in the block, though we must also account for additional off-diagonal terms in E due to the blocking of data points. Following the derivation steps above we can derive the β_{MAXb} for each block,

$$\beta_{MAX_B} = \sum_{i \in B} \frac{2\big(E(y^{(n)} - \mu)\big)_i}{\big(\sum_{j \notin B} 2w_j E_{j,i} + E_{i,i} + \sum_{k \in B} E_{k,i}\big)} \tag{4.8}$$

This can be used in a lightly modified version of Algorithm 5 where the β_{MAX_B} of blocks, not individual points, is iteratively computed.

Generalized Rayleigh Quotient method

We consider an alternative optimization approach to obtain an approximately optimal subset. Consider plugging the MLE solution, β^*, into w^* from Equation 4.2,

$$w^* = \underset{w}{\text{argmax}} \Big[w^T \big(E(y^{(n)}\text{-}\mu)(y^{(n)}\text{-}\mu)^T E\big) w\Big] / \Big[w^T \big(2E\big) w\Big] \tag{4.9}$$

If we relax w such that $w \in \mathbb{R}^m$, this can be re-written as the generalized Rayleigh quotient,

$$R(A, B, w) = \frac{w^T A w}{w^T B w}, \tag{4.10}$$

where $A = E(y^{(n)} - \mu)(y^{(n)} - \mu)^T E$, and $B = 2E$. Note that A is a symmetric matrix and B is a Hermitian positive-definite matrix. Taking the Cholesky decomposition $B = LL^T$, the generalized Rayleigh quotient can be written as a Rayleigh quotient [160], $R(A', w') = (w'^T A' w')/(w'^T w')$, where $A' = L^{-1}AL^{T^{-1}}$ and $w' = L^T w$. The maximum w' of the Rayleigh quotient, $w'_{max} = \text{argmax}_{w'} R(A', w') = \text{argmax}_{w'}(w'^T A w')/(w'^T w') = v^{(max)}$, is the largest eigenvector of A'. Since we defined $w' = L^T w$, then the maximum $w_{max} = L^{T^{-1}} v^{(max)}$ is the relaxed solution to our original optimization problem from Equation 4.9.

Although w_{max} has non-integer elements, the ordering of the elements of this eigenvector corresponds to the importance of the data points in the neighborhood. Thus we scan over the ordered elements of w_{max}, iteratively adding each to the subset. Maximizing $LLR(w)$ over this linear number of subsets provides an approximate solution to the constrained integer program.

Forward stepwise optimization

A third approximation approach uses a greedy forward stepwise algorithm that iteratively sets one element $w_i = 1$ such that the objective is minimized in each iteration. Once the objective cannot be further minimized the optimization is terminated, thereby providing a greedy optimal solution. For a neighborhood of size k, the stepwise approach may require up to k iterations, evaluating $O(k)$ subsets at each iteration for a total of $O(k^2)$ computations.

4.3.3 Efficient Multi-Stream Search

Often we are interested in searching for anomalous patterns across multiple dimensions, or streams, of data. For example, anomalous patterns of damaged trees and sewer flooding can help localize severe storm damage. Multi-stream search can enhance the signal of subtle anomalies that affect multiple streams, and reduce false positive detections when perturbations in a single stream are not important to the application.

In principle, GPNS and GPSS can handle multiple streams by stacking the data from each stream and adding a final dimension to indicate from which stream the data came. Yet naive GP inference requires $O(n^3)$ complexity, so repeatedly concatenating data from multiple streams quickly leads to scalability issues. On the other hand, Kronecker-based scalability require a kernel that is multiplicatively decomposable over the input dimensions [127]. This implies that the prior correlation structure is the same over all data dimensions except for the stream indicator. For example, Kronecker structure in spatiotemporal settings constrains streams to have the same prior spatiotemporal correlations. This assumption is overly restrictive for the complex data in which we are interested.

Instead, we learn independent GPs for each stream of data and then scan over neighborhoods in the data jointly for all streams. Posteriors for each stream are independently inferred from the associated GP. Thus for streams $s = 1, ..., S$, the posterior distribution for subset scanning contains a block diagonal covariance,

$$\mathcal{N}\left(\begin{bmatrix}\mu_1\\ \vdots \\ \mu_S\end{bmatrix}, \begin{bmatrix}\Sigma_1 & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & \Sigma_S\end{bmatrix}\right)$$

In this manner each stream can flexibly learn different prior covariance structures while still ensuring scalability equivalent to single-stream GPNS and GPSS. The one drawback of this approach is that inter-stream covariance information is not exploited for GP inference.

4.4 Experiments

We evaluate GPNS and GPSS using numeric simulations and three urban spatiotemporal datasets. We compare the methods against a number of competitive baseline algorithms from contemporary literature. First, we compare to an independent Gaussian subset scan, a state of the art anomalous pattern detection Algorithm [101, 102]. Additionally, we compare against a standard GP anomaly detection approach [84, 140], in which we use the posterior distribution of the null GP model θ_0 regressed over the entire dataset to classify points beyond a given level-α significance threshold as anomalies. While all GP methods in this chapter are agnostic to kernel choice, an RBF kernel and linear mean function were used for all experiments.

Although anomalous pattern detection is a distinct problem from outlier or anomalous point detection, we also compare against two commonly used outlier detection techniques: a one-class SVM [129] and robust multivariate outlier detection using the Mahalanobis distance [123, 124].

4.4.1 Numeric experiments

For each numeric test, baseline data is drawn from a two-dimensional GP [116]. Multiplicative anomalies of arbitrary shape are injected by scaling randomly sampled points, within a randomly chosen neighborhood, by a factor of ≥ 1. (Note that this simulation does not correspond to our method's assumption of an additive mean shift.) The most anomalous subset is computed using GPSS methods and baseline approaches. For the baseline GP approach and one-class SVM we provide additional information (the true percentage of the

anomalous data) in order to determine their threshold levels. Thus those baseline methods have more information than the GPSS approaches.

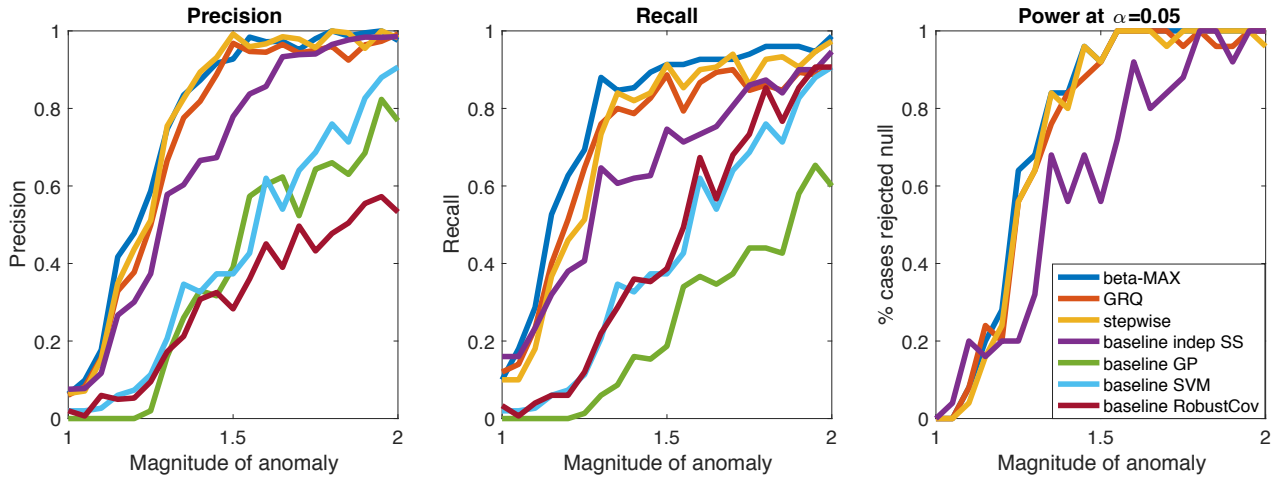


Fig. 4.1 Precision, recall, and power at $\alpha = 0.05$ for GPSS methods and baseline anomaly detection approaches. The three GPSS methods dominate in all cases with the β_{MAX} performing best overall.

Varying the multiplicative factor between 1 and 2 we compute the average precision and recall in Figure 4.1 over 50 tests in a 400 point grid for each multiplicative factor. Randomization testing ($\alpha = .05$) is performed for each synthetic test to determine the score threshold for significance. For precision and recall, truly anomalous points are "positive" and all other data is "negative." The GPSS approaches dominate all other methods for nearly the entire test range, with β_{MAX} performing best overall.

Additionally, for each test we use an exhaustive search to find the subset with the highest LLR. The ratios of the LLR of approximate GPSS solutions to $LLR(w^*)$ are shown in Figure 4.2. Note that all approximation methods are relatively close to the optimal value. While the β_{MAX} approach dominates at large magnitudes, the GRQ dominates at small magnitudes and achieves a relatively stable ratio across all tests. Such stability may be valuable when considering unexplored data.

To test the methods' scalability we vary the maximum neighborhood size and measure run time. In Figure 4.2 we compare GPSS, GPNS, and an exhaustive search for the optimal subset. The exhaustive search quickly becomes computationally intractable. Despite the added flexibility, GPSS is faster than GPNS because GP posterior inference is performed for fewer neighborhoods.

We consider the effect of the density of anomalies on GPSS and GPNS where "density" is defined by the proportion of anomalous points in the true subset (Figure 4.3). High densities represent compact anomalies, while low densities represent irregularly-shaped

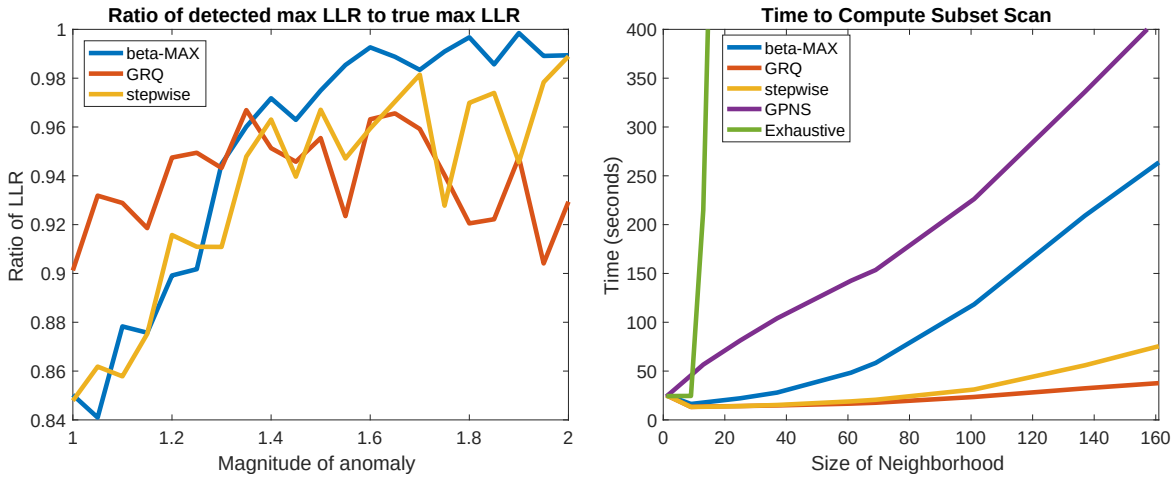


Fig. 4.2 Numeric tests of GPNS and GPSS compared to exhaustive evaluation of $LLR(w^*)$. Left plot: ratio of maximum LLR identified by GPSS to true maximum LLR. Right plot: run time.

anomalies. While the stepwise method is competitive with the β_{MAX} and GRQ approaches at low densities, its precision and recall drop off steeply at high densities. Thus the performance in of stepwise in Figure 4.1 would degrade significantly if we had chosen a higher density anomaly. Additionally, in relatively low density anomalies, where the anomalous shapes may be highly irregular, GPNS has substantially reduced precision and recall.

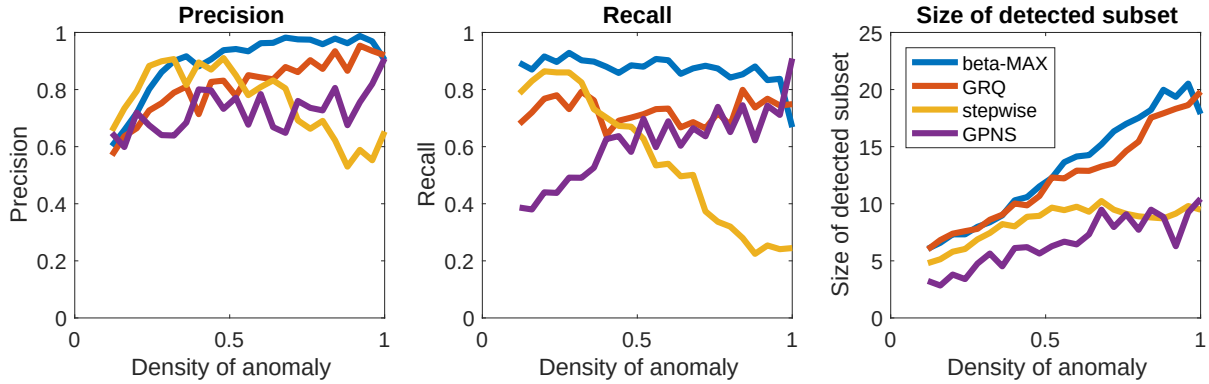


Fig. 4.3 Precision, recall, and size of detected subset for GPSS and GPNS methods over subsets of varying density within a neighborhood.

4.4.2 Urban opioid overdose deaths

A recent United States opioid epidemic has garnered national attention [145]. We study monthly opioid overdose deaths in New York from 1999-2015 [144]. Data is provided at

a county level for Manhattan, Brooklyn, Queens, the Bronx, Nassau County, and Suffolk County. Data is missing for some months in different counties. We apply GPSS and baseline approaches jointly to data across all time, latitude, and longitude, with randomization testing at $\alpha = 0.05$.

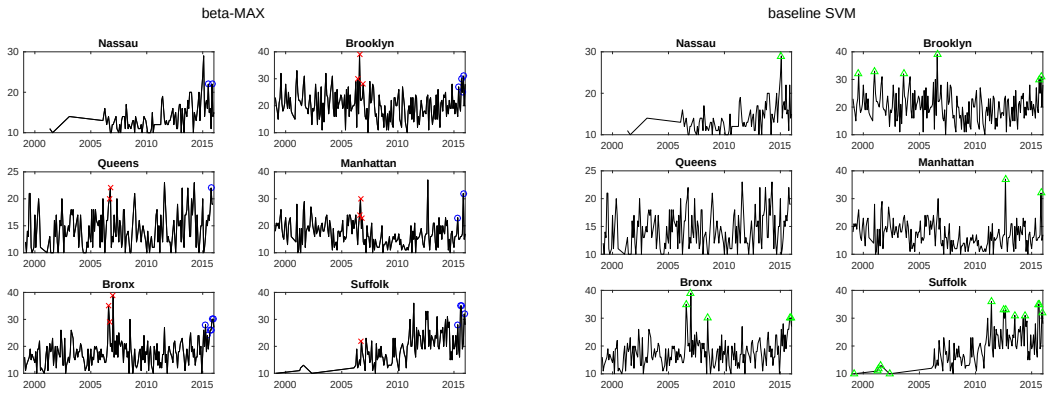


Fig. 4.4 Monthly opioid overdose deaths in New York from 1999-2015. Top plot depicts the two statistically significant anomalies detected by β_{MAX}. Bottom plot depicts points detected by the one-class SVM.

All three GPSS approaches (β_{MAX}, GRQ, and stepwise) identify two statistically significant anomalous patterns. While precise points selected by the methods differ slightly, Figure 4.4 depicts the two anomalous regions discovered by β_{MAX} in blue circles and red crosses. With the exception of the independent subset scan, the baseline methods failed to discover a coherent anomalous pattern. Instead they selected individual points across space and time. For example, see results from the one-class SVM in Figure 4.4.

The anomalies detected by GPSS correspond to important public health events. The blue circles at the end of 2015 indicate a surge in opioid deaths corresponding to a well known plague of fentanyl-related deaths in NYC [38]. The anomaly denoted by red crosses in 2006 is particularly interesting since it indicates a spike in opioid deaths immediately preceding the introduction of community training programs to administer a lifesaving naloxone drug. This may indicate a surge in fatalities that was cut short by making naloxone more widely available and educating communities in its use.

4.4.3 School Absenteeism

Public schools in New York City record and publish daily student attendance [107]. Given the importance of education on future outcomes there is tremendous interest in understanding

patterns of school absenteeism. We consider public school attendance data in Manhattan for the 2015-2016 school year. The data is messy, with missing entries and non-uniform placement of school locations. We aggregate data at weekly level and remove the last four weeks of the school year since they contain known high absenteeism rates that are not of interest to Department of Education officials.

We apply GPSS methods and baseline approaches with neighborhoods of up to ten local schools. All GPSS methods identified an anomaly around January to February 2016 concentrated on West Side of Manhattan. The results from GRQ around the time of the detected anomaly are presented in Figure 4.5. Each dot represents a school location, with yellow dots indicating high attendance and blue dots indicating low attendance. The space-time locations of schools in the top ten anomalous subsets are bordered in red.

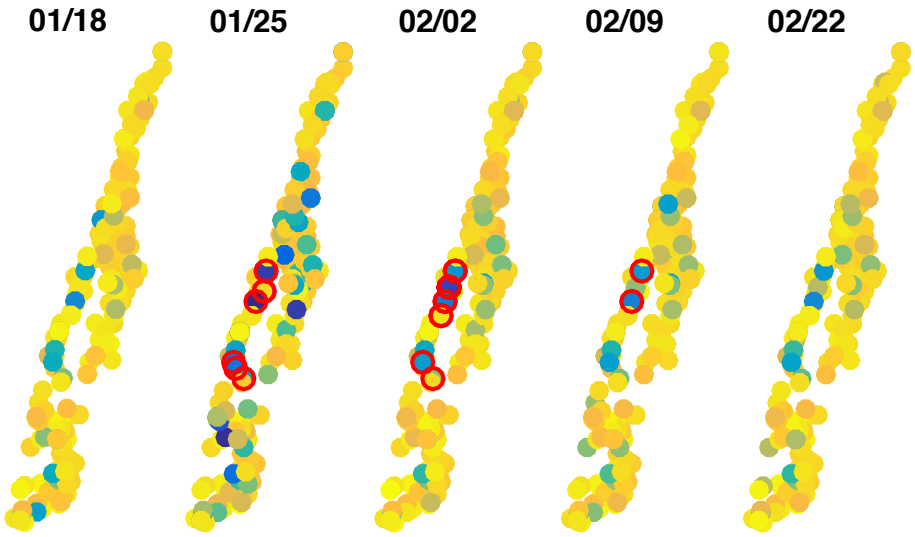


Fig. 4.5 School absenteeism results from Manhattan using GRQ. Each dot represents a school location, with yellow dots indicating high attendance and blue dots indicating low attendance. The space-time locations of schools in the top ten anomalous subsets are bordered in red.

The detected anomalies correspond to a category five blizzard which may have disrupted teachers and students from attending school even though no snow day closings were reported at the time. Further research is required to understand why the West Side of Manhattan differed systematically from the rest of the borough. Baseline anomaly detection methods did not identify a coherent anomaly and instead detected anomalies throughout the year.

4.4.4 Manhattan 311 requests

New York City's 311 system enables residents to request government services. We consider a local public health event that occurred on 01/22/16 in upper Manhattan. On that day, local news reported that residents were concerned due to brown tap water [109, 27]. Detecting the extent of the residents' concerns is important to help identify and mitigate public health risks.

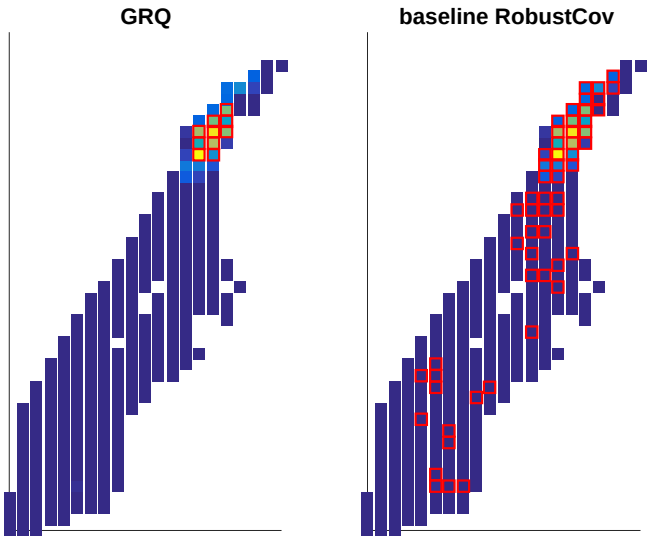


Fig. 4.6 GPSS and robust covariance results for daily 311 requests in Manhattan on 01/22/16. Red squares indicate detected anomalies.

We consider daily 311 requests in Manhattan for the month of January 2016, aggregated over a 0.08 mile^2 grid [37]. We apply GPSS methods and baseline approaches with neighborhoods of up to 15 points. All GPSS methods identified an anomalous pattern around the locations and time of the water discoloration event. Baseline methods tended to substantially overestimate the anomaly's extent in both space and time. These results from January 22 are represented by the GRQ and the Robust baselines in Figure 4.6. Blue and yellow squares indicate low and high volume of reports, respectively. Red squares indicate the top anomalous regions discovered by each method.

Ground truth does not exist for these hyper-local events so we cannot compute precision and recall. However, 311 requests have labeled types, although we used aggregated 311 calls as our data inputs. For each method we compute the ratio of water-related 311 calls to non-water-related calls in the detected anomalies. This "water signal-to-noise" ratio, listed in Table 4.1, indicates how precisely each method identified regions associated with many water-related requests. The entire dataset has a water signal-to-noise of 0.07.

Table 4.1 Signal-to-noise ratio of water-related 311 calls to non-water-related 311 calls for all methods.

Model	Signal-to-Noise
GRQ	7.22
Stepwise	7.22
β_{MAX}	7.22
Independent SS	7.06
Baseline GP	0.44
One-class SVM	0.23
RobustCov	0.12

4.4.5 Multi-stream: trees and sewers

Using the multi-stream procedure from Section 4.3.3, we consider 311 reports of damaged trees and sewer issues. Both streams indicate weather-related issues: damaged trees indicate high winds while sewer calls indicate substantial precipitation. Together, these data identify areas with dangerous post-storm conditions. Each complaint type is fit with an independent GP and the entire data is scanned jointly for anomalies.

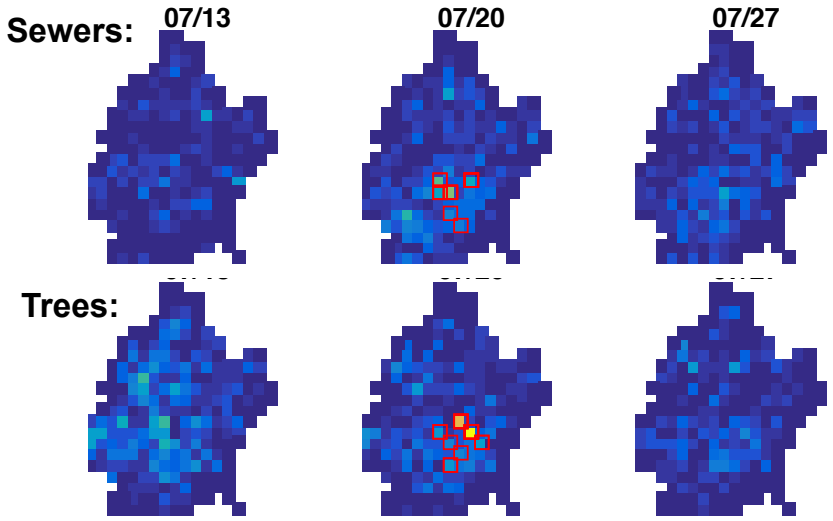


Fig. 4.7 311 calls for damaged trees and sewer issues from 2016 in Brooklyn. Red squares indicate the top anomalies discovered by the β_{max} approach.

We analyze data in Brooklyn aggregated weekly over a 0.08 mile^2 grid [37]. We conduct analyses for 2016 and 2010 with results depicted in Figs. 4.7 and 4.8. The number of sewer reports (per week, per cell) are plotted on top, and damaged tree reports on bottom. Red squares indicate the top anomalous regions discovered using the β_{max} approach. Note that

searches were computed over each entire year; we are only showing the time periods in which anomalies were discovered.

The most anomalous regions in 2016 were all concentrated during the week of July 20th when a significant summer storm felled trees and flooded sewers, thus jointly affecting both data streams [28]. Conversely, although the week of July 13th experienced elevated reports of felled trees no anomalous region is detected since there is no corresponding increase in sewer flooding. This demonstrates how multi-stream search may help to regulate GPSS.

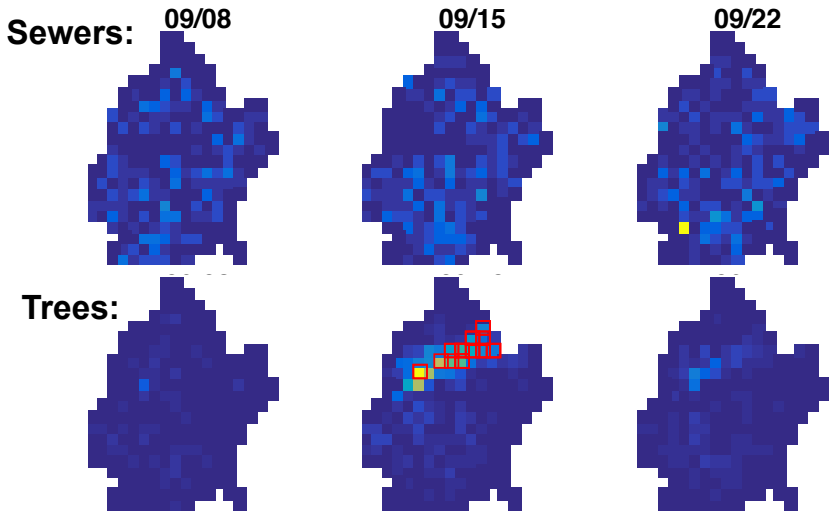


Fig. 4.8 311 calls for damaged trees and sewer issues from 2010 in Brooklyn. Red squares indicate the top anomalies discovered by the β_{max} approach.

The most anomalous regions in 2010 were all concentrated during the week of September 15th when an urban tornado cut through Brooklyn [5]. Unlike the 2016 results, these anomalies only occurred in reports of damaged trees. Also note the lone yellow square in the sewer data of September 22. Though the square indicates elevated number of calls, GPSS does not consider it anomalous since it does not represent a systematic shift in space and time.

Chapter 5

Regression Discontinuity Design Discovery

5.1 Introduction

Understanding causal mechanisms is critical for the social and laboratory sciences. While randomized control trials are the gold standard for identifying causal relationships, such experiments are often time consuming, costly, or ethically inappropriate. In order to exploit the plethora of observational data, econometricians often rely on "natural experiments," fortuitous circumstances of quasi-randomization that can be exploited for causal inference[1].

Regression discontinuity designs (RDDs) are such a technique. RDDs use sharp changes in treatment assignment for causal inference. For example, it is often difficult to assess the effect of academic interventions since treated students may systematically differ from other students. Yet, if a school intervenes on students who score below some threshold on a test, then students with scores just above or below the threshold are not systematically different and effectively receive random treatment [75]. That threshold induces an RDD that can be used to infer the effect of the intervention.

RDDs require fewer assumptions than most causal inference techniques and are arguably most similar to true randomized experiments [89]. However, identifying RDDs is a painstakingly manual process requiring human intuition and construction, and thus limited by human biases. Indeed, many papers reuse the same or analogous RDDs (e.g., discontinuities at geographic boundaries, or test score cutoffs for school admission) and most of these RDDs are one-dimensional, represented by a threshold value for a single variable. Finally, RDDs

[1] Published as Herlands et al. [62]

often rely on the human "eye" to verify their validity. The "tinkering" that is often done in practice implies that RDDs discovered by humans are subject to multiple testing issues.

To aid in discovering RDDs, we use statistical machine learning techniques to create the first general methodology to discover, quantify, and validate RDDs in data. Our approach can discover new RDDs across arbitrarily high dimensional spaces, enabling us to use RDDs that humans would not be able to identify otherwise. Yet these high dimensional RDDs are still interpretable, and we provide a simple mechanism for ranking how (observed) variables influence the discovered discontinuities. We derive two log likelihood ratio statistics to search for RDDs in potentially heteroskedastic data with either real-valued or binary treatments. Additionally, the technique can seamlessly handle both real-valued and categorical covariates. Finally, we present an integrated validation procedure ensuring rigorous statistical and econometric validity.

We evaluate our approach on synthetic and real data. Using synthetic data we demonstrate robust performance to out of sample discontinuities and model misspecification. For real data we consider three educational and health care settings previously studied in the econometric literature. Our approach can identify the RDDs in these data even with the injection of substantial additional noise.

While this is the first method we know of that discovers RDDs in general data, Card et al. [25] search for race-based "tipping" points in housing markets using an RDD design. They employ two search methods specific to the problem formulation: one inspired by the shape of curves derived from their data, and one that draws on structural break literature in time series [56]. Beyond RDDs, there is increased interest in integrating econometric and machine learning techniques [15, 100]. For example, deep learning and non-parametric Bayesian methods have been used to predict counterfactuals and compute individualized treatment effects [65, 79, 57]. Additionally, novel approaches have been developed for identifying heterogeneous treatment effects [71, 16]. Within the context of online recommendation systems, Sharma [135] and Sharma et al. [136] develop mechanisms of searching for certain natural experiments.

5.1.1 Outline

The remainder of the chapter proceeds as follows. Section 5.2 provides a brief overview of RDDs including their causal assumptions. Section 5.3 introduces our local search for RDDs including the search statistics used for the Normal (Section 5.3.1) and Bernoulli (Section 5.3.2) observation models, neighborhood definitions (Section 5.3.3), discontinuity validation (Section 5.3.4), and treatment effect estimation (Section 5.3.5). Section 5.4 discusses the synthetic and real data experiments.

5.2 Regression Discontinuity Designs

We provide practical background on RDDs for a computer science audience. There exist excellent papers for details on assumptions, inference, convergence, and model variations [55, 72, 146].

Throughout this chapter we consider data, (x,T,y), where $x = \{x_1,...,x_n\}$, are inputs that can include both categorical and real-valued $x_i \in \mathbb{R}^d$ variables, $T = \{T_1,...,T_n\}$ is a treatment variable that could either be binary, $T_i \in \{0,1\}$, or real-valued, $T_i \in \mathbb{R}$, and $y = \{y_1,...,y_n\}, y_i \in \mathbb{R}$, is an outcome variable. Both x and T are known *a priori* not to be affected by y. Additionally, we consider "forcing variables," z, which are a subset of the real-valued dimensions of x. Typically, the dimensions of $z \in x$ must be specified and validated by the user, but our algorithm does this automatically.

In the most straightforward RDDs, called "sharp RDDs," there is a one-dimensional forcing vector, z, and a cutoff value, c, such that before the cutoff value treatment is never assigned, $E[T|z < c] = 0$, and after the cutoff treatment is always assigned, $E[T|z > c] = 1$. Thus there is a sharp RDD at $z = c$ since at that point T jumps discontinuously from $T = 0$ to $T = 1$. As long as x does not also change discontinuously at $z = c$, there is no reason to believe that the data on either side of the discontinuity are systematically different. Thus, conceptually, at the local area around the discontinuity we can consider T to be randomly assigned. Notice that the RDD is a function of x, z, and T but not y. Indeed, an RDD allows us to investigate the effect of T on multiple different outputs, y.

RDDs appear in real-world settings where thresholds are used to assign treatment. For example, academic punishments given to students whose GPA drops below a specific value [91], or health insurance that covers children until they reach a certain age [11].

For this chapter we concentrate on "fuzzy RDDs" which generalize the sharp RDD. Fuzzy RDDs exist where T is partially determined by the discontinuity, i.e., where $P(T = 1)$ jumps discontinuously at $z = c$. The special case where that jump is from $P(T = 1) = 0$ to $P(T = 1) = 1$ constitutes a sharp RDD [72]. Given a fuzzy RDD, the treatment effect, τ, with respect to y, is,

$$\tau = \frac{\lim_{\varepsilon \to -0} E[y|z = c+\varepsilon] - \lim_{\varepsilon \to +0} E[y|z = c+\varepsilon]}{\lim_{\varepsilon \to -0} E[T|z = c+\varepsilon] - \lim_{\varepsilon \to +0} E[T|z = c+\varepsilon]}. \tag{5.1}$$

The limits in Eq. (5.1) indicate that although T is effectively random at $z = c$, farther away from the discontinuity T is not expected to be randomly assigned. That said, τ can be considered a weighted average treatment effect across the entire data, where the weights are ex-ante probabilities that a point is in the vicinity of $z = c$ [89].

The fuzzy RDD assumes the following conditions for identification [55] (the first two are also required for the sharp RDD):

- *Imprecise control*: the value of z cannot be precisely controlled to fall at $z = c \pm \varepsilon$. If such control did exist, those individuals manipulating z to be just above or just below c are likely to be systematically different than individuals who do not manipulate z, thus invalidating the design.
- *Excludability*: x crossing $z = c$ cannot affect y except through affecting the probability distribution of T.
- *Monotonicity*: x crossing $z = c$ cannot simultaneously cause some data to increase T and other data to decrease T.

These assumptions are relatively light and the first is even testable (see Section 5.3.4). Imprecise control replaces the ignorability or unconfoundedness assumptions that are necessary in many causal models. And unlike instrumental variables, RDDs do not assume anything about exogeneity [89]. Thus RDDs are quite suitable for automated discovery since they do not require the onerous, untestable, and often unbelievable assumptions made by other causal inference methods.

5.3 Method

The essential element of an RDD, which our approach aims to discover automatically from data, is the discontinuity, or "unexpected jump," in T. Given a model, $T_i = f(x_i) + \varepsilon_i$, this constitutes a special type of local anomaly where $f(x)$ substantially deviates from T both before and after the discontinuity. See Figure 5.1 for a 1-D example where $f(x)$ approximates the data well except for the two regions of deviation on either side of the discontinuity. Note that the deviations are of opposite sign and may be of different magnitudes. Traditional anomaly detection, such as one-class SVMs [129], focus on identifying individual outliers. Yet an RDD is fundamentally a pattern of multiple data points. Thus we employ anomalous *pattern* detection to search for RDDs. We frame the search as a log likelihood ratio (LLR) comparison between the likelihood of a null model that assumes no RDD exists, and the likelihood of an alternative model that assumes an RDD exists. We locally search for circumscribed neighborhoods that contain a discontinuity. Although any one neighborhood does not necessarily capture the entire discontinuity, it uses local data from around the discontinuity which can provide greater insight. The discovered discontinuities from multiple local neighborhoods can be combined to more precisely measure the treatment effect. Thus

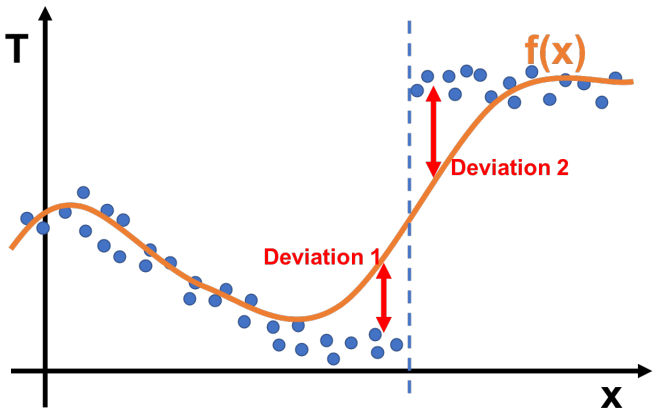


Fig. 5.1 Illustration of a one-dimensional RDD (dashed line). Blue dots are treatment T_i; orange line is $f(x_i)$.

our approach is named "**Lo**cal **R**egression **D**iscontinuity **D**esign **D**iscovery" (LoRD3). In a valid RDD, z must be be real-valued since sharp differences are expected to occur between data points with different values of a categorical variable. Thus, given data (x, T, y), we let all real-valued dimensions of x be forcing variables, z. LoRD3 searches for RDDs as follows[2]:

1. Model T with smooth model, $f(x)$, such that $T_i = f(x_i) + \varepsilon_i$.
2. Compute the estimated value of $\hat{T}$ using the learned model.
3. For each neighborhood size, $k = 1, \ldots, K$:
   (a) For each of the n data points, consider its k-sized neighborhood, $s_{i,k}$, defined by z (see Section 5.3.3).
      i. Compute the likelihood of a null model which assumes that $s_{i,k}$ does not contain an RDD: $L_0(s_{i,k})$.
      ii. Repeatedly bisect the neighborhood into two mutually exclusive partitions, assigning each point in the neighborhood to one of these two groups (see Section 5.3.3). We denote group assignment by $g_{k,j}$. For each grouping, compute the likelihood of an alternative model which assumes that $s_{i,k}$ contains an RDD with each group denoting one side of the discontinuity: $L_1(s_{i,k}, g_{k,j})$.

[2]Code and data are available at https://gitlab.com/herlands/LORD3

iii. Compute the maximum log likelihood ratio (LLR) over all partitions for that neighborhood,

$$LLR(s_{i,k}) = \max_j LLR(s_{i,k}, g_{k,j}) = \max_j \log \frac{L_1(s_{i,k}, g_{k,j})}{L_0(s_{i,k})}. \tag{5.2}$$

4. Test each of the neighborhoods for statistical significance and econometric validity, controlling for multiple hypothesis testing (see Section 5.3.4). For each "validated" neighborhood $s_{i,k}$ that passes these tests, record the corresponding $g_{k,j}$.

5. Estimate the τ using validated neighborhoods (see Section 5.3.5).

Notice that in step (a) the local neighborhoods are defined over the potentially multidimensional z. While most research using RDDs considers one-dimensional forcing variables z, even papers that consider multiple dimensional z [117, 158] require human identification and are limited in practice to low dimensions. LoRD3 seamlessly considers z of arbitrary dimension, allowing it to discover more diverse and nuanced RDDs than previously studied.

In Sections 5.3.1 and 5.3.2 we detail two observation models and LLR statistics for real-valued treatments and binary treatments respectively.

5.3.1 Normal residual observation model

Given a model, $T_i = f(x_i) + \varepsilon_i$, we would expect $f(x)$ to substantially and systematically deviate from a jump discontinuity in T. Specifically, near the discontinuity $f(x)$ should underestimate the true value of T on one side and overestimate T on the other side. We search for such a pattern using the LLR statistic below.

In principle we can use any regression approach for $f(x)$. Yet the appropriate choice requires $f(x)$ to be expressive enough to faithfully model the data, yet not substantially overfit to a potential discontinuity. For example, deep neural networks are untenable as they can model discrete jumps in data. In order to elucidate LoRD3, we consider polynomial models, $f(x) = \sum_{r=0:R} \gamma_r x^R$, which can be made increasingly expressive by increasing the polynomial order.

Given data (x, T), a neighborhood s, and a bisection g of the data points in s into "group 0" ($g_i = 0$) and "group 1" ($g_i = 1$), we consider the residuals, $r_i = T_i - f(x_i)$. The null model H_0 assumes that no discontinuity exists in s. We define the null model as r_i Normally distributed around a single offset parameter, β_0, which accounts for any bias in $f(x)$ over both groups. The alternative model H_1 states that a discontinuity exists between the two groups, and assumes that the r_i are Normally distributed around two distinct mean shifts, one

for each group. For maximal applicability, we consider unconstrained heteroskedastic noise, $\varepsilon_i \sim N(0, \sigma_i)$:

$$\begin{aligned} H_0 &: r_i \sim N\Big(\beta_0, \sigma_i\Big), \forall i \in s \\ H_1 &: r_i \sim N\Big((1-g_i)\beta_{g_0} + g_i\beta_{g_1}, \sigma_i\Big), \forall i \in s. \end{aligned} \tag{5.3}$$

Letting the alternative mean, $\mu_i = (1-g_i)\beta_{g_0} + g_i\beta_{g_1}$, for notational simplicity, we can compute the LLR,

$$\begin{aligned} LLR(s,g) &= \log \frac{Lik(H_1(s,g))}{Lik(H_0(s))} \\ &= \log\Big(\prod_{i\in s} P\big(r_i|N(\mu_i,\sigma_i)\big) \Big/ \prod_{i\in s} P\big(r_i|N(\beta_0,\sigma_i)\big)\Big) \\ &= \sum_{i\in s} (2r_i(\mu_i-\beta_0) - \mu_i^2 + \beta_0^2)/(2\sigma_i^2). \end{aligned} \tag{5.4}$$

For unrestricted heteroskedastic models we cannot directly compute σ_i. Instead, we assume that within a local area around each neighborhood the noise is homoskedastic. Thus we compute σ_i as the empirical variance in the k-neighborhood around each point.

We use the MLE values of β_0, β_{g_0}, and β_{g_1} from their respective heteroskedastic Normal models. Thus for each neighborhood,

$$\begin{aligned} \beta_0^* &= \Big(\sum_{i\in s} \frac{r_i}{\sigma_i^2}\Big) \Big/ \Big(\sum_{i\in s} \frac{1}{\sigma_i^2}\Big) \\ \beta_{g_0}^* &= \Big(\sum_{i\in s\cap(1-g)} \frac{r_i}{\sigma_i^2}\Big) \Big/ \Big(\sum_{i\in s\cap(1-g)} \frac{1}{\sigma_i^2}\Big) \\ \beta_{g_1}^* &= \Big(\sum_{i\in s\cap g} \frac{r_i}{\sigma_i^2}\Big) \Big/ \Big(\sum_{i\in s\cap g} \frac{1}{\sigma_i^2}\Big). \end{aligned} \tag{5.5}$$

5.3.2 Bernoulli log-odds observation model

For binary T, the Normal model is inappropriate since the residual between a binary variable and $f(x)$ is rarely Gaussian. Instead, we model T as a Bernoulli distributed random variable and search for discontinuities in the odds ratio [161].

Given a model for probability of treatment, $T_i \sim Bernoulli(p(x_i))$, we would expect $p(x)$ to systematically under- and over-estimate the true data around a jump discontinuity in T. We search for such a pattern using the LLR statistic below. We use a base model of a Logistic regression with polynomial functions, $p(x) = \text{Logit}(\sum_{r=0:R} \gamma_r x^r)$, to model the probability of a data point having $T_i = 1$.

Given data (x, T, s, g) as in Section 5.3.1, we consider the odds ratio of $T_i = 1$. The null model assumes that no discontinuity exists in s. We define the null model as a constant multiplicative scaled odds ratio to account for any bias in $p(x)$ over both groups,

$$H_0 : odds(T_i) = \beta_0 \frac{p(x_i)}{1 - p(x_i)}, \forall i \in s. \tag{5.6}$$

The alternative model assumes that a discontinuity exists between the two groups. Continuing to let $\mu_i = (1 - g_i)\beta_{g_0} + g_i\beta_{g_1}$, we define the alternative model as an odds ratio with two distinct multiplicative scales, one for each region,

$$H_1 : odds(T_i) = \mu_i \frac{p(x_i)}{1 - p(x_i)}, \forall i \in s. \tag{5.7}$$

These correspond to the null and alternative models,

$$\begin{aligned} H_0 &: T_i \sim Bernoulli\Big(\frac{\beta_0 p(x_i)}{1 - p(x_i) + \beta_0 p(x_i)}\Big), \forall i \in s \\ H_1 &: T_i \sim Bernoulli\Big(\frac{\mu_i p(x_i)}{1 - p(x_i) + \mu_i p(x_i)}\Big), \forall i \in s, \end{aligned} \tag{5.8}$$

with which we can compute the LLR,

$$\begin{aligned} LLR(s,g) &= \log \frac{\prod_{i \in s} P\Big(T_i | Bernoulli\big(\frac{\mu_i p(x_i)}{1 - p(x_i) + \mu_i p(x_i)}\big)\Big)}{\prod_{i \in s} P\Big(T_i | Bernoulli\big(\frac{\beta_0 p(x_i)}{1 - p(x_i) + \beta_0 p(x_i)}\big)\Big)} \\ &= \sum_{i \in s} T_i \log(\mu_i / \beta_0) + \log(1 - p(x_i) + \beta_0 p(x_i)) \\ &\qquad - \log(1 - p(x_i) + \mu_i p(x_i)). \end{aligned} \tag{5.9}$$

Unlike in the Normal case, there is no closed form solution for the MLE of β_0, β_{g_0}, or β_{g_1}. Instead, we solve for their values using a binary search. Eq. (5.10) provides the derivative of the log likelihood with respect to β_0. Note that the sum is taken over all points in the neighborhood ($i \in s$). Similar results hold for β_{g_0}, summing over group 0 ($i \in s \cap (1 - g)$), and β_{g_1}, summing over group 1 ($i \in s \cap g$).

$$\frac{\delta LL(s,g)}{\delta \beta_0} = \sum_{i \in s} \Big(\frac{T_i}{\beta_0} - \frac{p(x_i)}{1 - p(x_i) + \beta_0 p(x_i)} \Big) \tag{5.10}$$

We can then solve $\beta_0 \frac{\delta LL(s,g)}{\delta \beta_0} = 0$ by an efficient binary search, noting that this quantity decreases monotonically with $\beta_0 > 0$.

5.3.3 Neighborhood definition and bisection

Using only $z \in x$ to measure distance, the local neighborhood of a point includes itself and its $k-1$ nearest neighboring points. Since we are interested in generalizing to arbitrary dimensional RDDs we first compute the vector, $v_{s,i}$, between the center point of neighborhood s and each point $i \in s$. Then we bisect the neighborhood with $k-1$ hyperplanes, each of which passes through the center point, and is orthogonal to a $v_{s,i}$. Within each neighborhood, LoRD3 selects the bisection that maximizes the LLR defined above, testing the alternative hypothesis that there is an RDD for that neighborhood and bisection against the null hypothesis of no RDD.

5.3.4 Validate RDD neighborhoods

LoRD3 produces $O(n)$ neighborhoods - one centered at each data point - each with a corresponding bisection and $LLR(s)$. We can automatically assess which neighborhoods are statistically and econometrically valid using three techniques:

Randomization testing As is typical when searching with LLR statistics [85, 102], we use randomization to adjust for multiple testing and determine whether discontinuities are significant at level α. Specifically we use the following procedure:

1. Draw data $T^{(q)}$ from the null model Q times at the same covariates, x, as the true data.
    - In the Normal observation model, since $T_i = f(x_i) + \varepsilon_i$ this corresponds to sampling the noise Q times.
    - In the Bernoulli observation model, each T_i can be drawn directly from the H_0 Bernoulli distribution in Eq. (5.8).
2. Run LoRD3 on each $(x, T^{(q)})$. For each run save the value $LLR^{(q)} = \max_s LLR(s)$.
3. Compute an α threshold using the $1-\alpha$ quantile of the $LLR^{(q)}$ values. Any original neighborhoods s with $LLR(s)$ above this threshold are considered statistically significant.

For the unconstrained heteroskedastic model, we estimate each point's σ_i from the variance of data within the k-local neighborhood of that point, as in Section 5.3.1 above. Since LoRD3 evaluates $O(kn)$ possible neighborhood bisections, randomization is critical to address multiple testing issues. Since we use the maximum score over s for both original and replica data, this procedure provides an exact test for the highest-scoring neighborhood and a conservative test for secondary neighborhoods.

Density discontinuity As discussed in Section 5.2, RDDs assume that precise manipulation of T is not possible. A violation of this assumption could be reflected in a discontinuous density of z since data might "bunch" in z around the discontinuity to affect treatment status. McCrary [97] provides a commonly used procedure to test for such discontinuities in z. Since this test is limited to one dimension, we map our data to the vector orthogonal to the hyperplane that bisects the two groups in each neighborhood and apply the test on this one-dimensional data [42]. For each s, if the split selected by LoRD3 rejects the null we invalidate this s.

Placebo Testing In RDDs, placebo testing ensures that the discontinuity in T cannot be explained by a corresponding discontinuity or imbalance in x. While the forcing variables z are continuous within each neighborhood s, any $x \setminus z$, such as categorical variables, may still present issues. In order to be conservative, we run placebo tests on every dimension in x. We iteratively select one observational variable, $x^{(d)}$, and considering data $(T, x \setminus x^{(d)})$, we estimate τ with $x^{(d)}$ as the output (see Section 5.3.5 for how to compute $\hat{\tau}$). We ensure that this $\hat{\tau}$ is statistically indistinguishable from zero.

5.3.5 Estimating the treatment effect τ

Given a validated set of neighborhood discontinuities from LoRD3, practitioners may wish to further investigate the detected regions using domain expertise. Yet, it is also possible to directly use the neighborhood results from LoRD3 to estimate the treatment effect τ of treatment T on some real-valued output y. Below we describe three automated approaches for computing the estimate $\hat{\tau}$ given the results from LoRD3. If LoRD3 detects more than one validated neighborhood s, we compute $\hat{\tau}_s$ for each s and average them for the final estimation. Pooling the regions themselves, such as Bertanha [19] suggests for RDDs with multiple thresholds, is not possible in this case since there is no defined orientation of the two groups.

2SLS estimator A two-stage least squares estimation of τ first instruments $\hat{T}$ with a validated RDD neighborhood and then regresses $\hat{T}$ on y [13]. Given the data in neighborhood s, and indicators g_i for which group each data point is in, we first estimate,

$$T_i = \nu g_i + f(x_i) + \varepsilon_i^{(T)}. \tag{5.11}$$

Then we use the predicted $\hat{T}$ to regress (where λ is a learned vector),

$$y_i = \hat{\tau}\hat{T}_i + \lambda x_i + \varepsilon_i^{(y)}. \tag{5.12}$$

Non-parametric estimator Given a neighborhood and bisection, a non-parametric estimator for τ draws on Eq. (5.1). Assuming that the neighborhood is sufficiently small to approximate the limit, we use the empirical expectations over y and T to compute,

$$\hat{\tau} = \frac{E[y|g=1] - E[y|g=0]}{E[T|g=1] - E[T|g=0]}. \tag{5.13}$$

Group instrument While the 2SLS works generally for RDDs, using LoRD3 we can leverage information about μ to instrument T in each group. For the Normal model we instrument,

$$\hat{T}_i = T_i - \mu_i, \tag{5.14}$$

while for the Bernoulli model we can instrument $\hat{T}$ as,

$$\hat{T}_i = \frac{\mu_i p(x_i)}{1 - p(x_i) + \mu_i p(x_i)}. \tag{5.15}$$

Then we can run the second stage regression from Eq. (5.12).

5.3.6 Forcing variable influence

When humans identify an RDD it is clear which variables are responsible for the discontinuity. Since we consider potentially high dimensional z it is useful to identify which z variable(s) are most responsible for the RDD. Given a neighborhood, consider v_s, the vector orthogonal to the bisecting hyperplane. After normalizing the individual components of v_s to lie in $[0,1]$, those components indicate which dimensions of z most influence the discontinuity. For multiple neighborhoods, we average multiple normalized $v_1, \ldots, v_S$.

5.3.7 Evaluating discontinuities

Given a known discontinuity in synthetic or real data, we can evaluate how well a neighborhood s and bisection g chosen by LoRD3 correspond to the true discontinuity. Accuracy and precision are not appropriate metrics since there is no defined orientation of the two groups in a neighborhood. Instead, letting $d \in \{0,1\}$ define the space on either side of the true discontinuity, we compute the information gain (IG) of a k-sized neighborhood,

$$\begin{aligned} IG = k\,H\left(\frac{|s \cap d|}{k}\right) - |s \cap (1-g)|\,H\left(\frac{|s \cap (1-g) \cap d|}{|s \cap (1-g)|}\right) \\ - |s \cap g|\,H\left(\frac{|s \cap g \cap d|}{|s \cap g|}\right), \end{aligned} \tag{5.16}$$

where $H(p)$ is the entropy, $H(p) = -p\log(p) - (1-p)\log(1-p)$. We then normalize the IG to lie in $[0,1]$ by dividing by the optimal IG for a neighborhood of size k with a bisection of points into two equally sized groups, $k * H(\frac{1}{2})$. This metric is optimized when the neighborhood bisection overlaps fully with the true discontinuity and when the bisection equally divides the neighborhood points.

We provide measures of the normalized information gain (NIG) for all experiments in Section 5.4. Higher NIG is better since it indicates that a neighborhood bisection provides information about the true discontinuity. Lower NIG indicates that either the bisection is misaligned or the neighborhood does not intersect the discontinuity.

5.3.8 Practical considerations

As a pre-processing step before running LoRD3, we remove any data points with missing values and normalize each real-valued dimension x_j to have zero mean and unit variance. For datasets with categorical variables, we include these in x but not in z. Thus we do not consider heterogeneous treatment effects [70]. By default, all real-valued x are in z, though users may exclude variables based on domain knowledge. Finally, we note that approaches which analyze a known RDD may fit two background functions - one to each side of the discontinuity [89]. As detailed in Section 5.3, LoRD3 assumes a single background function in order to enable an efficient search. Additionally, we do not consider nonparametric $f(x)$ models such as local linear regression [72], but these could be easily incorporated.

5.4 Experiments

In order to demonstrate the power and flexibility of LoRD3, we apply the technique to a wide variety of synthetic and real data. RDDs are injected into the synthetic data, while for the real data we consider previously studied settings where known discontinuities exist. Note that the known RDD locations are used for evaluation purposes only, and are not provided to LoRD3. We inject additional noise into the real data to stress-test the search technique and evaluate its performance in the face of increasingly subtle discontinuities. For one-dimensional RDDs, we provide a comparison to existing changepoint detection methods in the literature.

5.4.1 Generating synthetic data

For synthetic experiments, we draw observed covariates, $x \in \mathbb{R}^d$, and unobserved covariates, $u \in \mathbb{R}^1$, through independent draws from a Uniform distribution, such that for $i = 1 \ldots n,\ j =$

$1 \dots d$,

$$x_{i,j} \sim \text{Uniform}(0,1), \quad u_i \sim \text{Uniform}(0,1). \tag{5.17}$$

We induce a discontinuity by randomly selecting a boundary, $b_j \sim \text{Uniform}(0,1)$ and defining an indicator,

$$D_i = \bigcup_{j=1}^{d} x_{i,j} > b_j. \tag{5.18}$$

Thus the discontinuous region is a d-dimensional cube and out-of-class for the hyperplanes LoRD3 uses to bisect each neighborhood. Throughout all experiments we consider heteroskedastic noise,

$$\varepsilon_i^{(T)}, \varepsilon_i^{(p)}, \varepsilon_i^{(y)} \sim N\left(0, \frac{1}{d}\sum_j x_{i,j}\right). \tag{5.19}$$

Real-valued treatment indicators T are generated by selecting the magnitude of the discontinuity, $\zeta \in \mathbb{R}$, and drawing,

$$\begin{aligned} \gamma_T &\sim N(0, I_d) \\ \mu_i &= I(x_i \in D)\frac{\zeta}{2} - I(x_i \notin D)\frac{\zeta}{2} \\ T_i &= x_i\gamma_T + \mu_i + \varepsilon_i^{(T)} + u_i. \end{aligned} \tag{5.20}$$

Binary treatment indicators T are generated by selecting the magnitude of the discontinuity, $\zeta > 0$, and drawing,

$$\begin{aligned} \gamma_p &\sim N(0, I_d) \\ \mu_i &= I(x_i \in D)\exp(\zeta/2) + I(x_i \notin D)\exp(-\zeta/2) \\ p_i &= \text{Logit}(x_i\gamma_p + \mu_i + \varepsilon_i^{(p)} + u_i) \\ T_i &\sim \text{Bernoulli}(p_i). \end{aligned} \tag{5.21}$$

Outputs $y_i \in \mathbb{R}$ are generated by selecting $\tau \in \mathbb{R}$ and drawing,

$$\begin{aligned} \gamma_y &\sim N(0, I_d) \\ y_i &= x_i\gamma_y + T_i\tau + \varepsilon_i^{(y)} + u_i. \end{aligned} \tag{5.22}$$

5.4.2 Synthetic real-valued treatment results

We generate real-valued T with $x \in \mathbb{R}^2$ and $\tau = 5$. To demonstrate how LoRD3 performs under different signal levels of discontinuity, we vary $\zeta \in [0, 2.5]$. For each ζ value we

generate 50 experiments with 1000 data points. For LoRD3 we let $k = 50$, $z = x$, and consider the top scoring neighborhood for evaluation. Base $f(x)$ models are order $r = 1, 2, 4$ polynomials to demonstrate results from both correctly and incorrectly specified models. Randomization testing is performed to determine an $\alpha = .05$ level for significance for each experiment. Finally, throughout the synthetic and real data experiments we have verified the placebo tests, as detailed in Section 5.3.4.

Figure 5.2 provides an example of an experiment with $\zeta = 3$. The axes are the dimensions of x. The left plot depicts T as colored circles and the square discontinuity is observable in the upper right. The center plot depicts $LLR(s)$ centered on each data point. The outline of the discontinuity has relatively high $LLR(s)$ indicating that LoRD3 has correctly identified neighborhoods along the boundary of the discontinuity. The right plot highlights the two groups from the neighborhood bisection with highest $LLR(s)$.

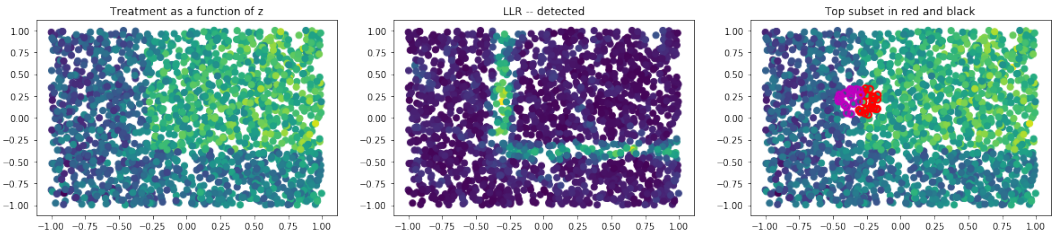


Fig. 5.2 Left plot: synthetic $T \in \mathbb{R}$ as a function of x. Center plot: $LLR(s)$ for each neighborhood using Normal model. Right plot: neighborhood bisection with highest $LLR(s)$.

We present results for NIG and power in Figure 5.3. LoRD3 performance improves as ζ increases since higher ζ induce a larger magnitude discontinuity. While the more complex specifications of $f(x)$ have slightly decreased performance due to overfitting, both NIG and power for all models are quite similar, demonstrating that the approach is robust to model misspecification.

Estimates of the treatment effect $\hat{\tau}$ are plotted in Figure 5.4. Due to the data generating process of y in Eq. (6.26), at low ζ where there is little to no discontinuity, LoRD3 tends to overestimate the true τ. However, all $f(x)$ model specifications (polynomials of degree 1,2,4) yield $\hat{\tau}$ that converge towards the true τ at larger ζ. While the non-parametric and 2SLS approaches converge more slowly, they tend to be more robust to model misspecification.

Varying dimension Letting $\zeta = 2$ and holding z fixed at two dimensions, we vary the number of covariates from 2 to 20. We apply LoRD3 with the three $f(x)$ models as above and plot the resulting NIG in the left panel of Figure 5.5. Next we hold x fixed at 10 dimensions and vary the number of dimensions in z from 1 to 10, plotting the results in the right panel of

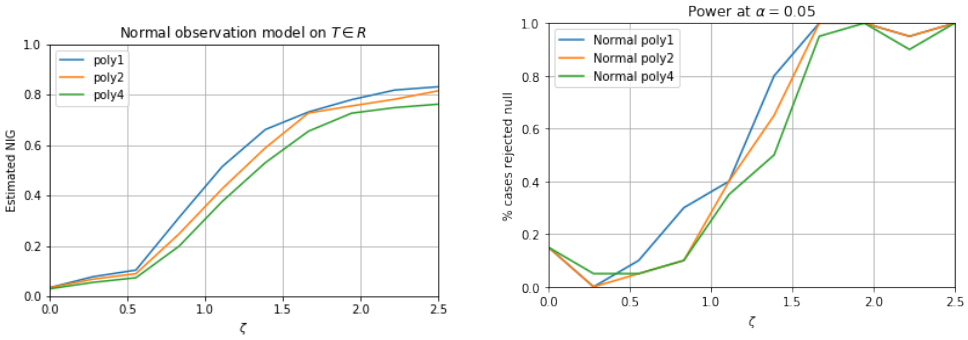


Fig. 5.3 Left: NIG of top neighborhood for $T \in \mathbb{R}$. The x-axis indicates ζ. Right: power to reject the null at $\alpha = 0.05$.

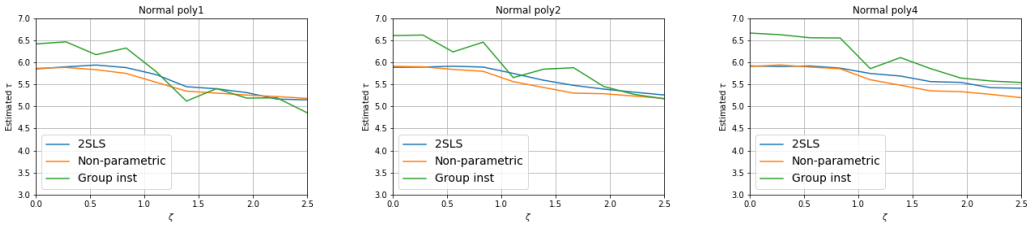


Fig. 5.4 LoRD3 Normal model estimated $\hat{\tau}$ on $T \in \mathbb{R}$. Each plot represents a different $f(x)$ specification. True $\tau = 5$.

Figure 5.5. These results indicate that given the same amount of data LoRD3 performance is robust to large numbers of covariates but reduces in performance over larger spaces of forcing variables.

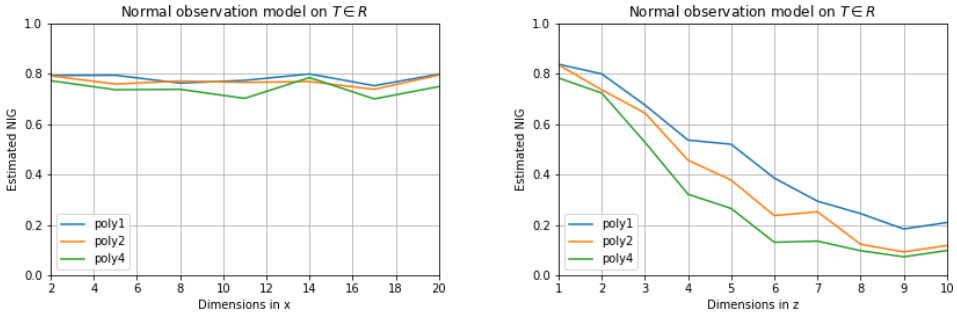


Fig. 5.5 NIG of LoRD3 Normal model for $T \in \mathbb{R}$ with varying the dimensions of x and z in left and right plots, respectively.

5.4.3 Synthetic binary treatment results

We generate equivalent synthetic tests for $T \in \{0,1\}$. For each experiment we run LoRD3 with both Normal and Bernoulli observation models. We use $p(x)$ of order $r = 1,2,4$ polynomials to demonstrate results from correctly and incorrectly specified models.

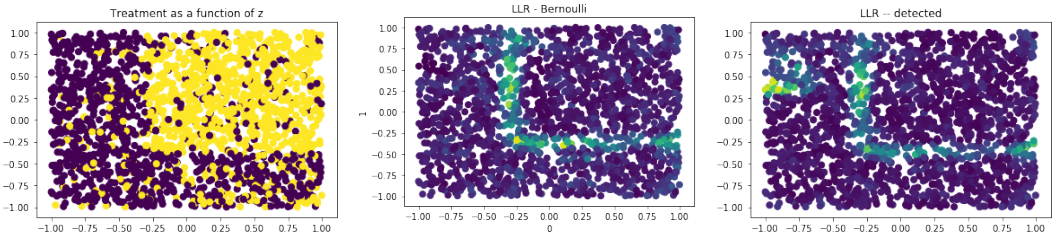


Fig. 5.6 Left shows $T \in \{0,1\}$ as a function of x, center shows $LLR(s)$ of Bernoulli model, $LLR(s)$ of Normal model.

Figure 5.6 provides an example of an experiment with $\zeta = 4$ and $LLR(s)$ using both the Bernoulli and Normal models. While both models discover neighborhoods with high LLR around the discontinuity boundary, the Normal model detects spuriously high LLR elsewhere in the space. The advantage of the Bernoulli model for binary T is also seen through the NIG results in Figure 5.7 where all $p(x)$ specifications using the Bernoulli model outperform the Normal model. We plot $\hat{\tau}$ from the Bernoulli model in Figure 5.8 where all $p(x)$ specifications have $\hat{\tau}$ that converge to the true $\tau = 5$ at larger ζ.

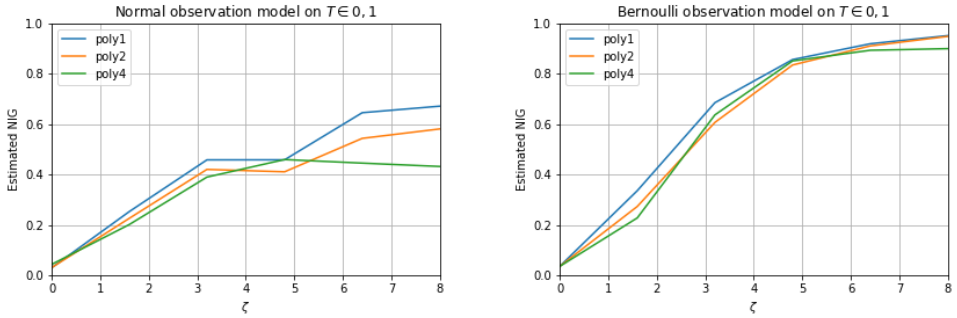


Fig. 5.7 NIG of top neighborhood for $T \in \{0,1\}$. Left plot: Normal model. Right plot: Bernoulli model.

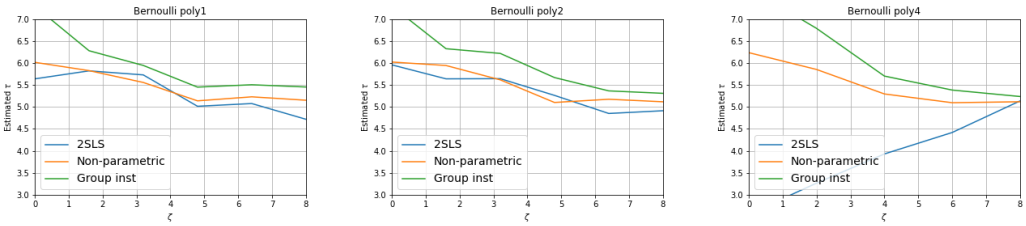


Fig. 5.8 LoRD3 Bernoulli model $\hat{\tau}$ on $T \in \{0,1\}$. Each plot represents a different $p(x)$ specification. True $\tau = 5$.

5.4.4 Comparison to changepoint detection

In one dimension, RDD discovery is similar to changepoint detection, where the objective is to identify points between regimes with persistent changes in mean or covariance structure. We consider competitive changepoint methods that utilize Binary Segmentation cluster analysis [132], parametric methods using Bartlett [60] and Student-t [59] test statistics, and non-parametric methods using Mann-Whitney [122] and Kolmogorov-Smirnov [121] test statistics.

We generate one-dimensional $T \in \mathbb{R}$ using $\zeta \in [0, 2.5]$ (see Section 5.4.1). For each ζ value we generate 50 experiments with 1000 data points. We apply all changepoint methods and LoRD3 with the Normal model and three $f(x)$ specifications. Mean squared error from the true discontinuity is used to evaluate the results in Figure 5.9.

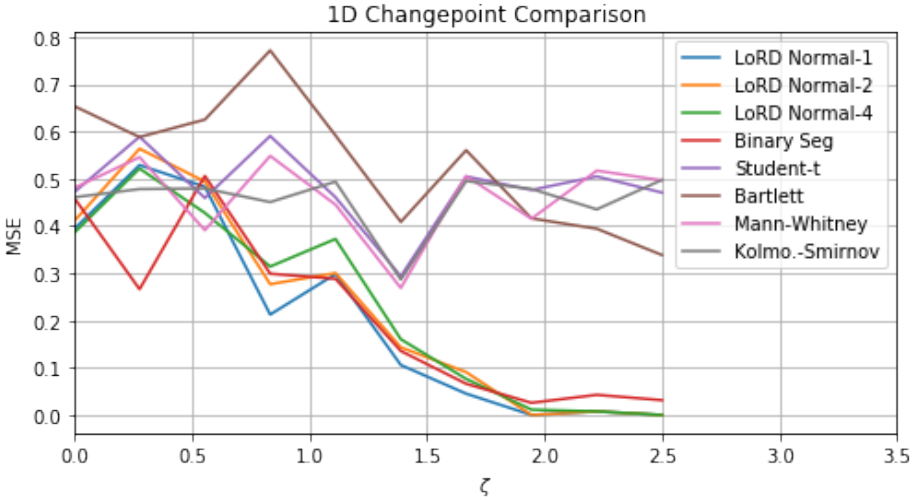


Fig. 5.9 Comparison of LoRD3 with changepoint methods.

We observe that all LoRD3 configurations are superior to changepoint methods for high ζ. Binary Segmentation equals the performance of LoRD3 MSE at low ζ, but has worse MSE than LoRD3 as ζ increases. Moreover, we note that these changepoint methods are limited to one dimension. LoRD3 advances into new territory by discovering RDDs in arbitrary dimensions and thus may be considered a generalization of changepoints to multiple dimensions.

5.4.5 Student test score data

Jacob et al. [75] consider the effect of an educational intervention on math test scores. We use their student test score dataset which is based on seventh-grade math assessments. It contains two sets of scores: "pre-test" scores that reflect student achievement before a potential intervention, and "post-test" scores after the intervention. Only students who received below 215 on the pre-test were intervened upon. Thus there is a sharp RDD at pre-test score 215.

The data has 2,606 observations and eight covariates, x, for each student. Six covariates are binary indicators for gender, special education status, eligibility for reduced-price lunch, English as second language status, and ethnicity (Black, White, Hispanic or Asian). Two of the covariates are real-valued: age of student and pre-test score. We use both real-valued variables as z even though only pre-test score is the true relevant variable. The intervention status is T and the post-test score is y. The true value of τ is 10.

We apply LoRD3 with the Normal and Bernoulli models, $k = 100$, and a 1-degree polynomial for $f(x)$ and $p(x)$. $LLR(s)$ is depicted in Figure 5.10. The strip of high LLR around pre-test score 215 indicates that LoRD3 was able to locate the discontinuity with both observation models.

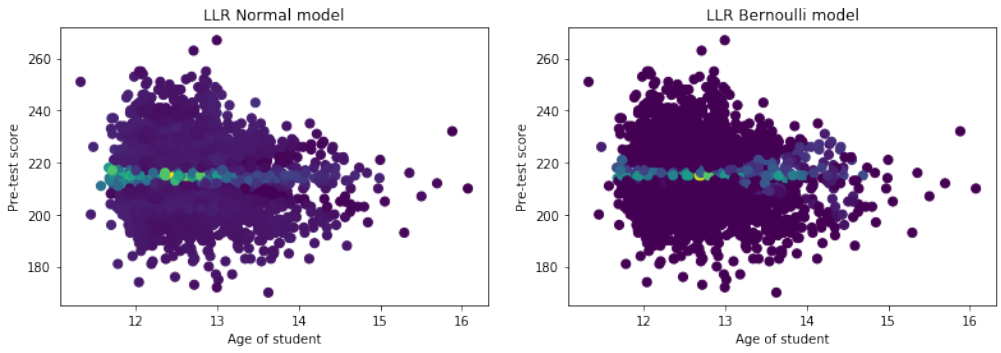


Fig. 5.10 $LLR(s)$ with student age as x-axis and pre-test score as y-axis. Normal model on left, Bernoulli model on right.

Table 5.1 lists the NIG, influence of the two z dimensions, and $\hat{\tau}$ over the top ten scoring neighborhoods. Both observation models yield high NIG and correctly identify pre-test score as the primary discontinuity variable. While the 2SLS and group instrument methods generally correctly yield $\hat{\tau}$ within the standard error of 10, the non-parametric method underestimates τ in both models.

Table 5.1 NIG, influence of z, and $\hat{\tau}$ for the student test data.

	Normal model	Bernoulli model
NIG	0.92 ± 0.02	0.93 ± 0.04
Influence: pre-test score	1.0 ± 0.0	1.0 ± 0.0
Influence: age of student	0.0 ± 0.0	0.0 ± 0.0
$\hat{\tau}$ 2SLS	8.89 ± 1.11	9.88 ± 0.98
$\hat{\tau}$ non-parametric	6.66	5.91
$\hat{\tau}$ Group inst	9.57 ± 1.18	6.30 ± 1.14

While the true data contains a sharp RDD, we inject synthetic noise to increase the difficulty of the search problem. We generate noisy treatment, T_ρ, such that $P(T_{\rho,i} = T_i) = \rho$, where $\rho \in [0.5, 1]$. Thus when $\rho = 1$, $T_\rho = T$, and the data contains a sharp RDD. When $\rho = 0.5$, $T_{\rho,i}$ is 0 or 1 with equal probability, resulting in no signal. Between those two extremes, the data exhibits a fuzzy RDD at pre-test score 215. Figure 5.11 depicts T_ρ at $\rho \in \{0.5, 0.75, 1\}$ to provide intuition for the magnitude of ρ noise.

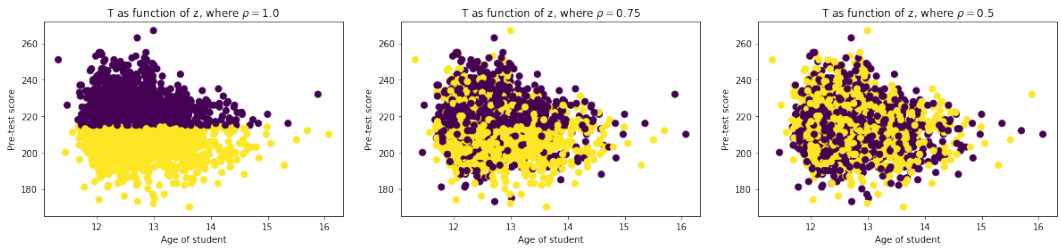


Fig. 5.11 Student data with pre-test score on y-axis, age of student on x-axis, and T indicated by circle color. Left plot is $\rho = 1$ (true T), center plot is $\rho = 0.75$, and right plot is $\rho = 0.5$.

For each value of $\rho \in [0.5, 1]$, we generate 25 experiments with 2000 randomly sampled data points. We apply LoRD3 as above and show results from the top scoring neighborhood in Figure 5.12. Both observation models converge to nearly $NIG = 1$ well before $\rho = 1$, demonstrating that they can identify RDDs in noisy data.

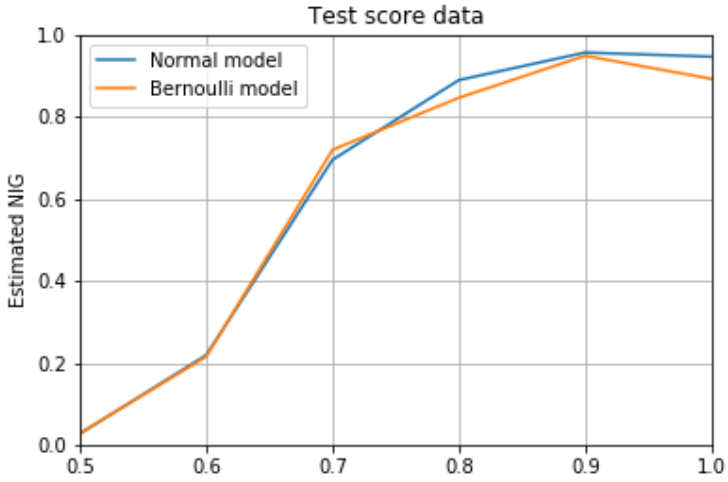


Fig. 5.12 NIG of top LoRD3 neighborhood on student test score data using Normal and Bernoulli observation models.

5.4.6 College GPA data

Lindo et al. [91] analyze the effect of academic probation on students at a Canadian university with three campuses. Students are placed on probation if their first year GPA is below a cutoff value. This cutoff induces an RDD that Lindo et al. [91] use to determine the causal effect of academic probation on educational outcomes.

The data has 44,362 observations and nine covariates, x, for each student. Five of the covariates are binary indicators for gender, English as a first language, being born in North America, and two variables to indicate which campus the student attended. Four covariates are real-valued: matriculation age, credits attempted in first year, high school grade percentile, and distance of GPA from the GPA cutoff. We use all four real-valued variables in z even though only distance from the GPA cutoff is the relevant factor. The intervention status is T. There are five outcomes of interest: decision to leave after the first academic term, GPA in the next academic term, and whether the student graduated within 4, 5 or 6 years.

We apply LoRD3 with Normal and Bernoulli models, $k = 100$, and $f(x)$ as a 1-degree polynomial. Table 5.2 lists the NIG and influence of the z dimensions over the top ten scoring neighborhoods. The Bernoulli model yields substantially higher NIG than the Normal model, as expected in data with binary T. Both methods correctly rank GPA cutoff as the most influential dimension of z, but the importance of this variable is less pronounced in the Normal results. Although [91] estimates $\hat{\tau}$ for each outcome using all students within 0.6 grade points of the cutoff GPA, these are not ground truth. Table 5.3 provides these values as

Table 5.2 NIG and influence of z for full university GPA data.

	Normal model	Bernoulli model
NIG	0.59 ± 0.05	0.71 ± 0.06
Influence: GPA cutoff	0.79 ± 0.37	1.0 ± 0.0
Influence: HS grade pct	0.59 ± 0.36	0.11 ± 0.13
Influence: credits yr 1	0.20 ± 0.40	0.0 ± 0.0
Influence: age of student	0.0 ± 0.0	0.0 ± 0.0

Table 5.3 Estimated $\hat{\tau}$ on university GPA data.

	Leave	GPA y2	Grad y4	Grad y5	Grad y6
Lindo et al. [91]	0.018	0.233	-0.020	-0.044	-0.024
Normal LoRD3 model					
2SLS	0.066	0.255	-0.211	-0.207	-0.116
Non-para	0.030	-0.025	-0.056	-0.189	-0.172
Group inst	0.058	0.188	-0.198	-0.187	-0.094
Bernoulli LoRD3 model					
2SLS	0.021	0.219	-0.273	-0.245	-0.050
Non-para	0.017	0.125	-0.076	-0.036	0.105
Group inst	0.020	0.055	-0.293	-0.098	0.144

well as LoRD3 $\hat{\tau}$ values using the methods in Section 5.3.5. Though we expect deviations between these estimates, most values, and nearly all signs, are quite similar.

In order to increase the difficulty of detection, we inject increasingly high ρ noise, as described in Section 5.4.5. For each ρ value we generate 25 experiments with 2000 randomly sampled data points. We apply LoRD3 with the same parameters as above and show NIG results for the top scoring neighborhood in Figure 5.13. In this case, there is substantial improvement using the Bernoulli model. While both models improve at higher values of ρ the Bernoulli increases to NIG$= 0.8$ while the Normal model only reaches NIG$= 0.7$.

5.4.7 Emergency department usage

Emergency department (ED) overcrowding and extended waiting times has been critical issue in the United States healthcare system [142, 69]. We consider aggregate emergency department (ED) patient data used to study the impact of health insurance on ED usage [10, 11]. Data come from 2.2 million ED visits between 2002-2009 in Arizona, California, Iowa, New Jersey, and Wisconsin. The only covariate is patient age and previous studies identified RDDs at ages 19 and 23. The existence of multiple discontinuities in this data is particularly interesting and we apply LoRD3 to see which RDDs it can detect. Note that due

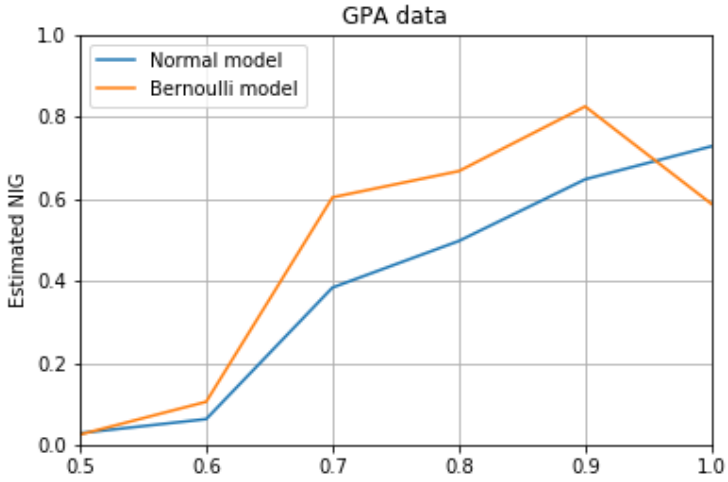


Fig. 5.13 NIG of top LoRD3 neighborhood on university GPA data using Normal and Bernoulli observation models.

to endogeneity issues Anderson et al. [10] develop a specialized τ estimation approach that is not replicated here.

Letting $x = z$ be ED patient age, we separately consider T as percentage of ED patients with private insurance and T as percentage of ED patients without insurance. In both cases we use $f(x)$ as a 3-degree polynomial and run 1000 randomization tests. We depict data, $LLR(s)$, and the $\alpha = 0.05$ significance threshold in Figure 5.14.

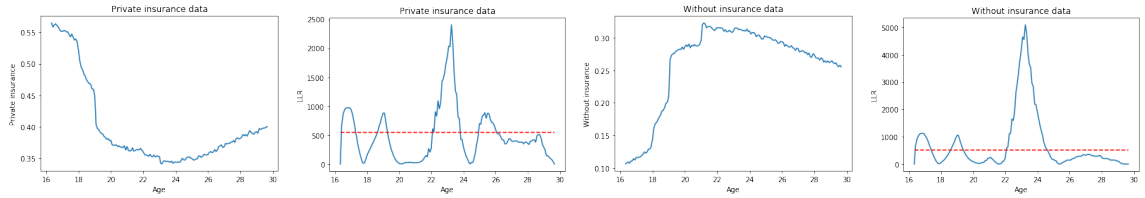


Fig. 5.14 ED patients with private insurance on top, without insurance on bottom. Left: % of patients vs. age. Right: $LLR(s)$ centered at each age. Red line indicates $\alpha = 0.05$ level.

The most prominent RDD peaks at 23 years 3 months for both T. This corresponds to the RDD used in Anderson et al. [11] and reflects that health insurance plans at the time allowed full-time students to remain on their parents' plans until age 23. Both setups also identify an RDD at age 19 corresponding to the RDD used in Anderson et al. [10]. This reflects that non-students were allowed to remain on their parents' insurance plans until age 19. Interestingly, both setups also identify an additional RDD centered at 16 years 10 months

Table 5.4 LoRD3 and changepoint comparisons for ED data.

	Private insurance	Without Insurance
LoRD3	16.83, 19, 23.25, 25.33	16.83, 19, 23.25
Binary Seg	18, 19.08, 22, 26.67	19.08, 21.08, 24.42, 27.08
Student-t	17.42	17.42
Bartlett	17.42	17.33
Mann-Whitney	16.92	17.25
Kolmo.-Smirnov	16.92	17.25

which may provide useful information for research. Finally, the setup with private insurance as T identifies a weaker RDD that peaks at 25 years 4 months. The identification of both known and unexplored discontinuities confirms the ability of LoRD3 to identify RDDs and to provide potentially policy-relevant insights.

We compare these results to the changepoint methods from Section 5.4.4. Binary Segmentation, which can find multiple changepoints, correctly identified the discontinuity at age 19, but was not able to discern the discontinuity at age 23. The remainder of the methods seem to corroborate that there is a discontinuity around age 17, though the precise values they detect differ slightly from LoRD3.

Chapter 6

Difference-in-Differences Discovery

6.1 Introduction

Understanding the effect of public policy interventions, such as new or proposed legislation, is critical for advancing impactful, evidence-based policies. However, laws are almost never enacted as RCTs. Instead, researchers must consider observational data before and after the policy implementation to understand its effects. In order to draw causal results from temporal observational data researchers often employ another natural experiment technique called difference-in-differences (DD).

DDs compare how a treated subpopulation and a "parallel", untreated control population change over time. In so doing, they address a key shortfall in RDDs. In particular, RDDs are not well suited for the analysis of changes occurring over time – something known as an RDD in time (RDiT). In RDiTs the forcing variable, z, is time, $z = t$. Since RDDs provide weighted average treatment effects around the treatment discontinuity, an RDiT can only provide insight about a treatment around the time of the discontinuity. The resulting treatment effect is bound to a particular time and may not be generalizable to the past or future. Indeed, time is a unique covariate insofar as we expect many changes to occur over time. Thus the treatment effect estimates from Section 5.3.5 may not be applicable to other times or over the longer term [58]. In contrast, DD are particularly built to analyze point-in-time policy changes. By considering cross sectional effects across two subsets—one treated, one untreated—DD enable us to generalize beyond the particular temporal moment of change.

Unfortunately, the discovery of DD is just as manual and serendipitous as finding RDDs. DD research often begins by considering a previously known intervention, such as new legislation, and analyzing if it induced a DD in observational data. The discovery process is further complicated since the identification of a parallel control subset is difficult and fundamentally based on unverifiable assumptions.

Identifying an appropriate control is a general problem in DD research. At its core, this question touches on the more general issue of determining an appropriate counterfactual. While RCTs are designed to have both treatment and control units, natural experiments do not have a pre-selected control. In many DD applications the control is created by selecting a subset of non-treated units and arguing that they represent an appropriate comparison for the treated units. This control subset may consist of all non-treated elements or even just a single non-treated element [43]. Part of the justification is based on demonstrating that the treatment and control subsets exhibit parallel trends in the pre-treatment period [12]. This implies that the counterfactual treatment and control would also be parallel in the post-treatment period a implied by the DD treatment effect estimation in Equation 6.4. However, parallel trends in the pre-treatment period does not guarentee counterfactual similarity. Further, recent work has shown that selecting a control based on tests for non-rejection of the parallel trends assumption may be misguided [22, 112]. Thus it is important that the control also be believably similar to the treatment subset. This is often achieved through qualitative argument.

Instead of selecting a subset of records to function as a control, another popular approach is to construct a "synthetic control" that satisfies certain conditions [4, 2]. One popular approach for synthetic controls is to learn a convex combination of non-treated units which, when combined in aggregate, exhibit parallel trends in the pre-treatment period. While this may address the concern about finding an appropriate subset with strongly parallel pre-trends, it may lead to substantial overfitting. There are deep connections between traditional DD setups and synthetic controls and [43] provides a useful framework unifying the two approaches.

To automatically discover DDs we modify the RDD subset scanning methodology from Chapter 5 in a two key ways. First we develop an extension of LoRD3 that can identify heterogeneous RDDs in categorical data. We apply this to data where the forcing variable is time, $z = t$, thus enabling us to discover RDiTs. Second, we develop a method for identifying a control subset parallel to the treatment subset. By pairing an RDiT with a parallel control, we can identify a treatment effect through standard DD calculations. Both of these developments represent novel methodological contributions and may be used independently. Specifically, the control identification may be used in nearly any DD application. However, when these methods are used together, they represent a self-contained automated search technique for DDs.

We evaluate the automatic DD search technique with synthetic and real data. Using synthetic data, we demonstrate robust performance to varying specifications and out of sample data. For real data we consider two policy settings where state-level legislation

induced a DD between states in the United States. Our approach can identify the DDs in these data and provide insight for public policy experts.

6.1.1 Outline

The remainder of the chapter proceeds as follows. Section 6.2 provides a brief overview of DDs including their causal assumptions. Section 6.3.1 introduces our search for heterogeneous RDiTs including multiple approaches for searching over heterogeneous subsets. Section 6.3.2 introduces our search for an appropriate control subset along with two comparison techniques. Section 6.4 discusses the synthetic and real data experiments.

6.2 Difference-in-Differences

We provide practical background on DD for a computer science audience. There exist excellent papers for details on assumptions, inference, convergence, and model variations including [13, 20, 12], among others.

Throughout this chapter we consider a dataset, D, that contains n records, $R_1, ...R_i, ...R_n$. An individual record is defined by $R_i = (x_i, t_i, \theta_i, y_i)$ where $x_i \in \mathbb{C}^m$ is a vector of categorical inputs such that each dimension $x_{i,j} \in \{v_{j,1}, ..., v_{j,k_j}\}$, $t_i \in \mathbb{R}$ are time, θ_i are treatments that could either be binary, $\theta_i \in \{0,1\}$, or real-valued, $\theta_i \in \mathbb{R}$, and y_i are outcomes that could either be binary, $y_i \in \{0,1\}$, or real-valued, $y_i \in \mathbb{R}$. τ is the treatment effect. We use the potential outcomes framework [126] where $y_i(0)$ denotes an outcome without treatment and $y_i(1)$ denotes an outcomes with a treatment. Since we never observe both $y_i(0)$ and $y_i(1)$ for any particular R_i, we must estimate one of the counterfactuals $\hat{y}_i(0)$ or $\hat{y}_i(1)$ in order to compute a treatment effect.

DDs are a popular econometric framework for determining the causal effect of an intervention in temporal observation data. In particular, DDs are used when there exist possible biases over time and between the treatment and control subsets. The most basic DD model estimates a constant additive intervention effect between two subsets and two time periods where the intervention, $\theta = 1$, exists only in the treatment subset at only one time period [14]. If we assume constant "fixed effects" for each subset and time period, letting $x_i = 1$ be the treatment subset and $x_i = 0$ be the control subset, this estimate can be viewed as a *difference of the difference* of expectations,

$$\tau = \Big[\mathbb{E}[y_i|x_i = 1, t_i = 1] - \mathbb{E}[y_i|x_i = 1, t_i = 0]\Big] \tag{6.1}$$

$$-\Big[\mathbb{E}[y_i|x_i = 0, t_i = 1] - \mathbb{E}[y_i|x_i = 0, t_i = 0]\Big] \tag{6.2}$$

Alternatively, the additive effect can be derived by a simple linear regression.

$$\begin{aligned} y_i &= \beta_0 + (\beta_t * t_i) + (\beta_x * x_i) + (\tau * \theta_i) + \varepsilon_i \\ \theta_i &= x_i * t_i \end{aligned} \tag{6.3}$$

Extending the problem to multiple control units, $x \in 1, ..., g$, and multiple time periods, $t \in 1, ..., h$, and continuing to assume constant fixed effects, we immediately arrive at the "panel DD" model which can also be solved via linear regression,

$$y_i = \beta_0 + \sum_{T=1}^{h} (\beta_T * \mathbb{1}_{t_i=T}) + \sum_{X=1}^{g} (\beta_X * \mathbb{1}_{x_i=X}) + (\tau * \theta_i) + \varepsilon_i \tag{6.4}$$

where $\theta_i = 1$ in a subset of treated units after the time of intervention. Using a DD or panel DD, we can estimate τ and its standard error. This estimation can be efficiently computed due to the simplicity of linear regression.

6.2.1 Assumptions

DDs must satisfy the assumptions required for general OLS regressions, including SUTVA. Additionally, at the boundary of the intervention, DDs must satisfy the same imprecise control assumption as RDDs.

Unique to DDs is the assumption of parallel trends. The primary purpose of parallel trends is to ensure that in the absence of treatment, the gap between the treatment and control subsets in the post-treatment period would be the same as it is in the pre-treatment period. If this is true, then the counterfactual subsets will be "parallel" across time. However, since we do not have access to this counterfactual world the assumption is unverifiable. Instead, we must measure how parallel the two subsets are in the pre-treatment period. Qualitative reasoning and the human eye are then often used as a general guide to ensure believability of the parallel trends assumption for the entire time including the treatment period.

6.3 Method

DDs are employed to analyze data across time, while RDDs in time (RDiTs) are generally avoided. Yet DDs and RDDs are intimately connected since both rely on sudden changes in θ. Indeed, DDs are composed of an RDiT in some subset of the data and another parallel untreated subset of the data. Thus, it is natural to consider adapting some of the LoRD3

methodology for DD discovery. When doing so there are two important differences between the natural experimental techniques that make searching for DDs more difficult:

- In RDDs, the θ discontinuity affects all cross sectional units at the boundary. In DDs, the discontinuous treatment only affects a subset of the cross sections over time. Thus a DD search needs to identify heterogeneous treated cross-sectional subsets that do not contain the entire dataset.
- After identifying the RDiT, to construct a DD, we then need an additional step of identifying an appropriate control subset of untreated cross sectional units that satisfies the parallel trend assumption.

With these considerations in mind we develop the Subset DD Discovery System (SuDDDS). This method shares the core concept of LoRD3: identifying discontinuities in treatment can be framed as an anomalous pattern detection problem. Yet SuDDDS requires developing substantial additional technology for the DD setting. Section 6.3.1 extends the LoRD3 methodology to identify heterogenous RDiT subsets in categorical data. Then Section 6.3.2 develops a novel approach for control subset identification that exhibit parallel trends within the paradigm of subset selection.

6.3.1 Heterogeneous RDiT Search

Consider a subset $s \in D$ to contain one or more covariate profiles comprising cross-sectional units across all points in time. We search over categorical subsets to discover a subset, s_τ, which contains a treatment discontinuity in time. Were we interested in identifying heterogeneous RDiTs, this step would be sufficient. Indeed, this step already represents a generalization of LoRD3 to a heterogeneous search over categorical subsets of data. While we concentrate on the special case where the forcing variable $z = t$, our approach can be trivially extended for any arbitrary z.

We define a log likelihood ratio (LLR) statistic between a null model, $H_0(s, T_0, W)$, which assumes that s does not contain a RDiT in the time window $T_0 - W < t \leq T_0 + W$, and an alternative model, $H_1(s, T_0, W)$, which assumes that s contains an RDiT during that time window,

$$LLR(s, T_0, W) = log \frac{L_1(s, T_0, W)}{L_0(s, T_0, W)} \tag{6.5}$$

where T_0 is the time of the intervention or treatment.

Algorithm 6 presents the top-level conditional optimization we employ to identify a heterogeneous RDiT. At the core of the algorithm we alternatively condition on T_0 and

optimize s_τ, and then condition on s_τ and optimize T_0. This conditional optimization is randomly initialized ℓ times in order to provide the algorithm with the opportunity to more fully explore the search space. In practice, for the first iteration we often initialize $s = D$ to consider a maximal subset consisting of the entire dataset.

Algorithm 6 Heterogeneous RDiT Search

1: **for** $W = W_1 : W_w$ **do**
2: **for** iteration $= 1 : \ell$ **do**
3: Initialize s_τ and T_0 randomly
4: **repeat**
5: Compute $LLR_{prev} = LLR(s_\tau, T_0, W)$
6: $T_0 = \max_T LLR(s_\tau, T, W)$ by Algorithm 7
7: $s_\tau = \max_{s \in D} LLR(s, T_0, W)$ by Algorithm 8
8: **until** $LLR(s_\tau, T_0, W) = LLR_{prev}$
9: **end for**
10: **end for**
11: Test s_τ for statistical significance and econometric validity

Conditionally optimize T_0 As noted in line 8 of Algorithm 6, given s_τ we conditionally optimize T_0. We approach this optimization through an exhaustive search detailed in Algorithm 7. At each step we bisect s_τ at some point in time into two mutually exclusive partitions and compute the *LLR* for each bisection. The time that induces a partition with the greatest *LLR* is returned as the new T_0.

Algorithm 7 Conditionally optimize T_0

1: Get the set of unique time points, $\mathbb{T}$, in s_τ
2: Compute $LLR(s_\tau, T, W)$ for each $T \in \mathbb{T}$
3: **return** $T_0 = \text{argmax}_T LLR(s_\tau, T, W)$

Conditionally optimize s As noted in line 9 of Algorithm 6, given T_0 we conditionally optimize s_τ. The mechanism for doing so is difficult. On the one hand, an exhaustive search over categorical subsets to identify an optimal s_τ would require an exponentially complex search in the number of data records. On the other hand, were we to naively select all the individual records that have individually high *LLR* when $t = T_0$, we would overfit s_τ to noise across a scattered subset of records. Instead we want the selected s_τ to represent a coherent set of records amenable to a DD estimation.

In order to compute this search in polynomial time we adapt the Multidimensional Subset Scan (MDSS), which has been previously employed for subset scanning over categorical

search spaces [103, 161]. Our approach to MDSS is detailed in Algorithm 8. Each iteration of the algorithm optimizes s_τ over the dimensions of x in a random sequence, which ensures that at each iteration we update s_τ such that $LLR(s_\tau, T_0, W)$ weakly increases. Note that

Algorithm 8 MDSS

1: **repeat**
2: Compute $LLR_{prev} = LLR(s_\tau, T_0, W)$
3: Randomly order the m dimensions of x to scan from 1 to M
4: **for** $j = 1 : M$ **do**
5: **for** $k = 1 : k_j$ **do**
6: $s_{j,k} = s_\tau \cap \{x_i | x_{i,j} = v_{j,k}\}$
7: Compute priorities $\gamma(s_{j,k}, T_0)$
8: **end for**
9: Scan over $s_{j,k}$ ordered by $\gamma(s_{j,k}, T_0)$.
10: Update s_τ with the highest scoring subset.
11: **end for**
12: **until** $LLR(s_\tau, T_0) = LLR_{prev}$
13: **return** s_τ

MDSS is a search over categorical values. For real-valued covariates of x we can discretize them into a pre-specified number of units. Additional covariates in x that are not intended to be included in the subset profile can still be used to model the treatment in line 1 of Algorithm 6.

Compute LLR: Double β Normal Residual Model

For the *LLR*, we consider a variant of the Normal residual observation model from Section 5.3.1 that detects the pair of deviations between the data and a smooth model immediately before and after a sharp change in treatment. As in Section 5.3.1 we employ polynomial models, $f(x) = \sum_{r=0:R} \gamma_r x^R$, which can be made increasingly expressive by increasing the polynomial order. After modeling the data, we compute residuals, $r_i = \theta_i - f(x_i, t_i)$, and assume they are Normally distributed such that,

$$\begin{aligned} H_0 &: r_i \sim N\Big(\beta_0, \sigma_i\Big), \forall i \in s \\ H_1 &: r_i \sim N\Big((1-g_i)\beta_{g_0} + g_i\beta_{g_1}, \sigma_i\Big), \forall i \in s. \end{aligned} \tag{6.6}$$

where g_0 is an indicator of the set of records occurring in time interval $[t-W, t)$ and g_1 is an indicator of the set of records occurring in time interval $[t, t+W]$. Letting $\mu_i = (1-g_i)\beta_{g_0} + g_i\beta_{g_1}$ be the alternative mean for notational simplicity, we can compute the

LLR,

$$\begin{aligned} LLR(s,g) &= \log \frac{Lik(H_1(s,g))}{Lik(H_0(s))} \\ &= \log \Big(\prod_{i \in s} P\big(r_i | N(\mu_i, \sigma_i)\big) \Big/ \prod_{i \in s} P\big(r_i | N(\beta_0, \sigma_i)\big) \Big) \\ &= \sum_{i \in s} (2r_i(\mu_i - \beta_0) - \mu_i^2 + \beta_0^2)/(2\sigma_i^2). \end{aligned} \tag{6.7}$$

MDSS requires a priority function, $\gamma(s,t)$, to rank order subsets of data (see Section 4.1 for a discussion of priority functions in subset scanning algorithms). Technically there are two sets of priority functions that we can construct for DDs: one for a discontinuity that occurs "forward" in time (i.e. where an event causes some s to discontinuously change in the future) and one that occurs "backward" in time. Due to the natural ordering of time the "forward" case is more intuitive and reflects how DDs are generally constructed in the literature. While all of the priority functions detailed in this chapter take the "forward" form, symmetric "backwards" functions would work as well.

For the priority function we let $c_i = \sum \frac{r_i}{\sigma_i^2}, b_i = \sum \frac{1}{\sigma_i^2}$ for each record and consider the aggregate terms,

$$\begin{aligned} C_1 &= \sum_g c_i \\ B_1 &= \sum_g b_i \\ C_2 &= \sum_{1-g} c_i \\ B_2 &= \sum_{1-g} b_i \end{aligned} \tag{6.8}$$

The *LLR* can then be reformulated as a function of these aggregated terms,

$$LLR(s,g) = \frac{C_1^2}{2B_1} + \frac{C_2^2}{2B_2} - \frac{(C_1 + C_2)^2}{2(B_1 + B_2)} \tag{6.9}$$

Following a similar approach to Neill [102] we define the ratios,

$$\begin{aligned} q_1 &= \frac{C_1}{B_1} \\ q_2 &= \frac{C_2}{B_2} \end{aligned} \tag{6.10}$$

where q_1 and q_2 are also the MLE values of β_{g_0} and β_{g_1}, respectively. This allows us to write the *LLR* as,

$$LLR(s,g) = HM(B_1,B_2)\Big(\frac{(q_1-q_2)}{2}\Big)^2 \tag{6.11}$$

where HM is the harmonic mean, $HM(\alpha,\beta) = \dfrac{2\alpha\beta}{\alpha+\beta}$. This formulation provides the motivation to rank slices of data by the difference between their q ratios,

$$\begin{aligned}\gamma(s,t) &= q_1 - q_2 \\ &= \frac{\sum_g c_i}{\sum_g b_i} - \frac{\sum_{1-g} c_i}{\sum_{1-g} b_i}\end{aligned} \tag{6.12}$$

Using this priority function, we *could* naively order records by (q_1-q_2) and then scan over that list, iteratively adding one record at a time to determine an optimal *LLR*. This approach intuitively makes sense since B_1 and B_2 (and thus their harmonic means) weakly increase as we increase the number of records in a subset.

However, the two q values can suffer from a version of Simpson's paradox, where the combined (q_1-q_2) may not be a convex combination of the original q_1 and q_2 values. Indeed, this priority ordering can provide substantially sub-optimal results in some non-convex cases such as when some high-priority elements have $q_1 >> 0$ and $q_2 = -\varepsilon$, and some have $q_1 = \varepsilon$ and $q_2 << 0$. In order to address this issue we consider the two methods below.

Greedy Search Instead of scanning directly over a list of records priority ranked by (q_1-q_2), we can follow a greedy search where we iteratively add to the subset the data element that maximizes the total combined (q_1-q_2). Using this priority method ranking we apply Algorithm 8.

While this greedy search resolves Simpson's paradox, it has other drawbacks. For example, consider that the top element has $q_1 >> 0$ and $q_2 = -\varepsilon$ and a number of other elements have $q_1 = \varepsilon$ and $q_2 << 0$. The greedy search will add only the top element, but an optimal search would forgo the top element in favor of a subset of the other elements. Therefore, while the greedy method will provide a better subset than a naive approach, it cannot guarantee an optimal subset.

Weighted Convex Combinations An alternative method is to reframe the (q_1-q_2) priority function as $(\rho)(q_1)+(1-\rho)(-q_2)$ where $0 <= \rho <= 1$. This enables us to consider any convex combination of the q values. Furthermore, within Algorithm 8 we can draw multiple values of ρ uniformly on $[0,1]$ and maximize over subsets from all of these priority functions. We can thus (separately) consider elements with $q_1 >> 0$ and $q_2 = -\varepsilon$ and those with $q_1 = \varepsilon$

and $q_2 << 0$, for ρ close to 1 and ρ close to 0 respectively. Intuitively, this enables us to correctly favor one q more heavily and thus obviate both the Simpson's paradox and greedy search issues.

Single Δ Normal Residual Model

Instead of modeling the DD with two β parameters, we can consider a single Δ offset for both sides. Still assuming Normality of the residuals from Section 6.3.1, we can model the null and alternative models,

$$\begin{aligned} H_0 &: r_j \sim N\Big(\mu_i, \sigma_j^2\Big), \forall j \in S_i \\ H_1 &: r_j \sim N\Big(\mu_i - g_0\Delta + g_1\Delta, \sigma_j^2\Big), \forall j \in S_i. \end{aligned} \tag{6.13}$$

where S_i represents a subset with the same covariate profile. Each covariate profile S_i has its own mean shift, μ_i, determined by maximum likelihood estimate (MLE) under H_0, and assumed to be identical under both H_0 and H_1. Under H_1 there is an additional shift of $\pm\Delta$ on opposite sides of the boundary, where Δ is constant across all covariate profiles S_i and determined by MLE.

Overall this single Δ Normal residual model represents a much more flexible model due to the covariate-profile level variables. Yet within a covariate profile the discontinuity model is slightly less expressive since we assume that the magnitude of the deviation between the data and $f(x,t)$ is the same before and after the treatment discontinuity. Importantly, this single variable in H_1 enables this model to avoid the difficult double q search in Section 6.3.1.

To derive an appropriate priority function, let,

$$\begin{aligned} B_i &= \sum_{R_j \in S_i} \frac{1}{\sigma_j^2} \\ C_i &= \sum_{R_j \in S_i} \frac{g_1(r_j - \mu_i) - g_0(r_j - \mu_i)}{\sigma_j^2} \\ C &= \sum_i C_i \\ B &= \sum_i B_i \end{aligned} \tag{6.14}$$

The *LLR* can then be reformulated as a function of these aggregate terms,

$$LLR(s,t) = \frac{-\Delta^2}{2} B + \Delta C \tag{6.15}$$

Thus the MLE $\Delta^* = \frac{C}{B}$ and we can use the priority function,

$$\gamma(s,t) = \frac{C_i}{B_i} \tag{6.16}$$

to rank the data. This enables efficient and exact computation of the priorities for the single Δ model.

Validating RDiT Subsets

As discussed in Section 5.3.4, we want to evaluate the RDiT to ensure statistical significance and econometric validity. For SuDDDS we employ the same three techniques as LoRD3, with slight variations.

Randomization testing Given the many subsets evaluated in SuDDDS we need to adjust for multiple hypothesis issues. We employ the same randomization testing procedure from Section 5.3.4 in order to ensure that the *LLR* of s_τ is statistically significant.

Density discontinuity The one-dimensional density discontinuity test from McCrary [97] is used to ensure there is no bunching at the discontinuity point. Unlike LoRD3, for SuDDDS there is no need to map data to single vector since $z = t$ is always unidimensional.

Placebo Testing Since $z = t$ we run placebo tests for each dimension of x. Using the control subset identification techniques from Section 6.3.2 we estimate $\hat{\tau}$ with one dimension of x as the output and ensure that $\hat{\tau}$ is statistically indistinguishable from zero.

6.3.2 Identifying a Control Subset

Once a heterogeneous RDiT has been found we need to identify an appropriate control subset in order to estimate a treatment effect, $\hat{\tau}$. It is important that control exhibit parallel trends with the treatment subset in the pre-treatment period. Yet, as discussed in Section 6.1 this does not guarantee an appropriate counterfactual. Therefore it is important that the control and treatment also be believably similar. In this section we consider two common approaches for control identification and then propose a novel method within the paradigm of subset selection.

Stated more technically, the objective is to estimate the counterfactual outcome $y_i(0) \; \forall \; \{i \in s_\tau; t_i \geq T_0\}$. By comparing this to the existing outcome data $y_i(1)$ we can compute the treatment effect. In order to identify an appropriate control subset we often try to minimize the

counterfactual MSE in the pre-treatment period,

$$\sum_{i \in s, t_i < T_0} \left(\hat{y}(0)_i - y_i \right)^2 \tag{6.17}$$

Standard DD Control

The first comparison method is a standard DD setup used in much of the literature. After defining the treatment subset, all other records, $D \setminus s_\tau$ are used to define the control. The counterfactual is computed from an average of the cross-sectional control records at each time point plus a constant offset,

$$\begin{aligned} \hat{y}(0)_i &= \alpha + \frac{1}{D \setminus s_\tau} \sum_{j \notin s_\tau, t_j = t_i} y_j \\ \alpha &= \frac{1}{T_0 |s_\tau|} \sum_{i \in s, t_i < T_0} y_i - \frac{1}{T_0 |D \setminus s_\tau|} \sum_{i \in D \setminus s_\tau, t_i < T_0} y_i \end{aligned} \tag{6.18}$$

The constant offset accounts for fixed effects between the treatment and control subsets (see Equation 6.4). This approach provides a believable control since, theoretically, by considering all non-treated units there is no manipulation or selection that could lead to overfitting in the pre-treatment period. However, this is not necessarily true in practice. Researchers have discretion about which records to include in their dataset so there may be effective manipulation of the control data by limiting the extent or scope of the dataset. Additionally, this approach for estimating $\hat{y}(0)_i$ may not result in parallel treatment and control groups, indicated by high pre-treatment MSE

Synthetic control

Another method of comparison is synthetic control, an approach that constructs a control by learning a convex combination of non-treated units. At its core the synthetic control estimates the counterfactual

$$\begin{aligned} y_i(0) &= \sum_{j \in D, t_j = t_i} w_j y_j \\ \sum_j w_j &= 1 \end{aligned} \tag{6.19}$$

We do not provide extensive details on the theory and additional constraints of synthetic controls which can instead be found in [2, 4, 9] among others. In practice we use the `Synth` R package to compute the synthetic control for all experiments [3].

Synthetic controls have the advantage of having lower pre-treatment MSE than the standard DD setup. However, this comes at the expense of potential overfitting. Additionally, since the weighting parameters, w are estimated based on observable data, the causal claims of the resulting treatment estimates may be limited [48, 81].

Greedy Expansion Subset Search (GESS)

We propose an alternative method for identifying a suitable control for s_τ. While the Standard DD framework provides a believable control, in a large dataset it is unlikely that simply considering all $D \setminus s$ will yield a reasonable estimate of the counterfactual to a relatively small treatment subset. An alternative approach is to search for the top-k records that minimize counterfactual MSE in the pre-treatment period [43]. Yet such an unconstrained search may also lead to overfitting; the resulting control subset may consist of records with unrelated covariate profiles – not a very believable or coherent control subset.

Instead, we want to balance the search for low counterfactual MSE with the intuition that the most similar records are those which share the most characteristics with s_τ. Those records are most likely to exhibit parallel post-treatment counterfactual outcomes. Additionally, the control subset should be compact without allowing for cherry-picking specific records. For example, if in a population study s_τ is white males, a control subset consisting of all males may have lower pre-treatment MSE than the entire population dataset. Additionally, even if a subset comprising only Asian females has lower MSE, such a group is less believably similar to s_τ than all males.

We search for this subset by greedily expanding s_τ to form a superset, s_{sup}, where the control subset is defined as $s_c = s_{sup} \setminus s_\tau$. The search optimizes s_{sup} such that s_c minimizes the pre-treatment MSE. We compute the counterfactual using a slightly modified version of Equation 6.18,

$$\begin{aligned} \hat{y}(0)_i &= \alpha + \frac{1}{|s_c|} \sum_{j \in s_c, t_j = t_i} y_j \\ \alpha &= \frac{1}{T_0 |s_\tau|} \sum_{i \in s_\tau, t_i < T_0} y_i - \frac{1}{T_0 |s_c|} \sum_{i \in s_c, t_i < T_0} y_i \end{aligned} \tag{6.20}$$

For notational purposes we let $mse(s_{sup})$ be the counterfactual MSE for the control subset defined by s_{sup}. We also define v_s to be the covariate profile of subset s. Thus $s = \{R_i | x_{i,j} \in v_s\}$. Our approach, detailed in Algorithm 9, extends s_{sup} by iteratively adding one covariate value to v_s. Each additional covariate value extends v_{sup} and thus expands s_c. This continues in a greedy manner until the counterfactual MSE declines.

We also consider a slightly altered version of GESS where instead of iteratively adding a single covariate's value to v_s we add all values from a covariate. To achieve this we replace

Algorithm 9 Greedy Expansion Subset Search (GESS)

1: Initialize $s_{sup} = s_\tau$
2: **loop**
3: **for** For each dimension, j of x **do**
4: Define $s^{j,k} = \{R_i | x_{i,j} \in (v_{s_{sup}} \cup v_{j,k})\}$ for each $v_{j,k} \in x_j$
5: **end for**
6: $s^{j,k}_{sup} = \text{argmax}_{j,k} mse(s_{j,k})$
7: **if** $mse(s^{j,k}_{sup}) < mse(s_{sup})$ **then**
8: $s_{sup} = s^{j,k}_{sup}$
9: **else**
10: Break
11: **end if**
12: **end loop**
13: **return** $s_c = s_{sup} \setminus s_\tau$

the definition of $s^{j,k}$ in line 4 of Algorithm 9 with,

$$s^{j,k} = \{R_i | x_{i,j} \in (v_{s_{sup}} \cup v_j)\}. \tag{6.21}$$

In a classic bias-variance tradeoff this alternative model reduces the flexibility of the model but also reduces overfitting in the pre-treatment period.

Randomization Testing

In order to ensure the statistical significance of $\hat{\tau}$ we employ randomization testing. We iteratively draw a subset, s_p, from a non-treated region of the data, $s_p \in \{D \setminus s_{tau}\}$ and compute $\hat{\tau_p}$. This provides us the ability to construct an empirical null distribution and test whether $\hat{\tau}$ is significant relative to this distribution. For example, for a result to be statistically significant at a level $\alpha = 0.05$ the observed value needs to lie above the 95th percentile of the null scores.

It is important to note this randomization test focused on $\hat{\tau}$ is independent of the procedures in Section 6.3.1 focused on the RDiT. We can separate out the significance testing of the treatment discontinues from the testing of outcome effects since our RDiT searching technique does not consider y. Indeed, the methods and testing from Section 6.3.2 can be applied to any DD scenario without regard to how the RDiT is initially identified.

6.4 Experiments

SuDDDS is evaluated using a rigorous combination of synthetic and real-world data. For the synthetic experiments we test the robustness of the algorithm by varying the discontinuity magnitude, treatment subset complexity, and treatment effect magnitude. We also consider the effect of heterogenous data generative processes in a single dataset. We further demonstrate the ability of SuDDDS to discover natural experiments in real world data. Two policy-relevant datasets contain data exhibiting complex behavior over multiple covariates. In both settings SuDDDS detects DDs corresponding to important policy changes. These applications demonstrate how the technique could provide insight to shape future policy.

6.4.1 Synthetic Data

To generate synthetic data we draw observed covariates, x, and unobserved covariates, u, by independent draws from a V-sized discrete uniform distribution, such that for $i = 1 \dots n,\ j = 1 \dots d$,

$$x_{i,j} \sim \text{DiscreteUniform}(1,V), \quad u_i \sim \text{DiscreteUniform}(1,V). \tag{6.22}$$

The data feature heteroskedastic noise,

$$\varepsilon_i^{(\theta)}, \varepsilon_i^{(p)}, \varepsilon_i^{(y)} \sim N\left(0, \frac{1}{d}\sum_j x_{i,j}\right). \tag{6.23}$$

We induce a treatment discontinuity by randomly selecting a subset, s_I, of the values in each of the dimensions of x and randomly selecting $T_0 \in t$. We draw γ_θ parameters for each categorical value in each dimension of x,

$$\gamma_{\theta,j,v} \sim N(0, I_{d*v}) \tag{6.24}$$

Real-valued treatment, θ, is then generated by selecting the magnitude of the discontinuity, $\zeta \in \mathbb{R}$, and drawing,

$$\theta_i = x_i \sum_{j \in d} \sum_{v \in V} \gamma_{\theta,j,v} I(x_j = v) + I(x_i \in s_I) I(t_i \geq T_0)\zeta + \varepsilon_i^{(\theta)} + u_i. \tag{6.25}$$

Outputs $y_i \in \mathbb{R}$ are generated by selecting treatment effect $\tau \in \mathbb{R}$ and drawing,

$$\begin{aligned} \gamma_y &\sim N(0, I_d) \\ y_i &= x_i \gamma_y + \theta_i \tau + \varepsilon_i^{(y)} + u_i. \end{aligned} \tag{6.26}$$

6.4.2 Synthetic Experiments

We generate categorical synthetic data with four dimensions of x, each with eight discrete values and ten time periods. The treated subset is defined by two random values in each of two randomly chosen dimensions. We vary the magnitude of the discontinuity and apply the RDiT heterogeneous search portion of SuDDDS from Section 6.3.1. Specifically we test both the Greedy and Weighted Convex methods for optimizing *LLR* in the double β Normal residual model from Section 6.3.1 as well as the single Δ Normal residual model from Section 6.3.1. Results for precision, recall, and F-score of the identified RDiT are shown in Figure 6.1. The Weighted Convex and Single Δ models perform similarly, while the Greedy approach has substantially lower recall than the other techniques. This reflects a sub-optimality of the greedy approach: that it may reach a local maximum and terminate before adding all treated units to s_τ.

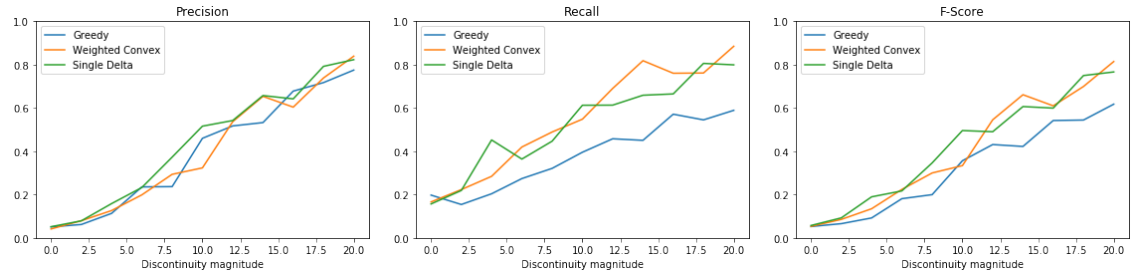


Fig. 6.1 Precision, recall, and F-score of RDiTs identified by SuDDDS at varying magnitudes of the true RDiT discontinuity. Three methods for optimizing *LLR* are compared.

Using a similar experimental setup, we vary the complexity of the true subset by adjusting the number of dimensions that define the intervention from 1 to 5 dimensions. We maintain a constant magnitude of the discontinuity at $\tau = 10$. Results for precision, recall, and F-score of identifying the RDiT are shown in Figure 6.2. The flat lines in these plots indicate that all three methods are quite robust to changes in the complexity of the intervention. This is important since in a real-world application the complexity of the intervention is not known a priori.

Using synthetic data we apply the three methods for identifying control subsets from Section 6.3.2, assuming that the true RDiT has been correctly identified in the first step of SuDDDS. We then use the the counterfactual output, $\hat{y}(0)$, from each method to compute an estimated treatment effect, $\hat{\tau}$. Finally, we compute the MSE between$\hat{\tau}$ and the true synthetic τ and plot the results in Figure 6.3. The left figure plots $\hat{\tau}$ of each control estimation method along with a dashed red line indicating the true τ. The right depicts the mean MSE of the individual treatment effect estimates. Under these data all methods correctly estimate τ.

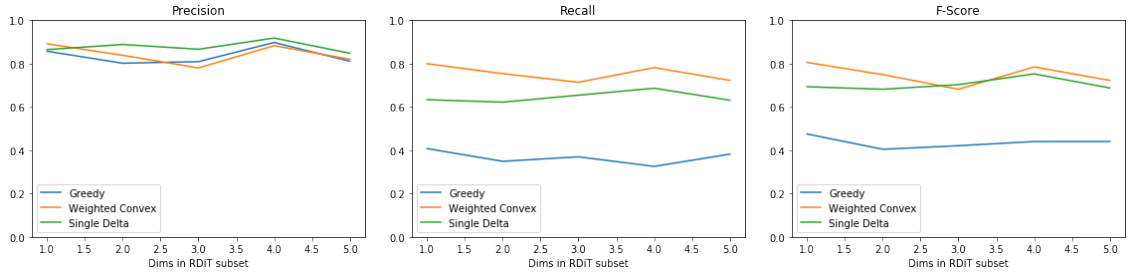


Fig. 6.2 Precision, recall, and F-score of RDiTs identified by SuDDDS at varying complexities of the true RDiT subset. Three methods for optimizing *LLR* are compared.

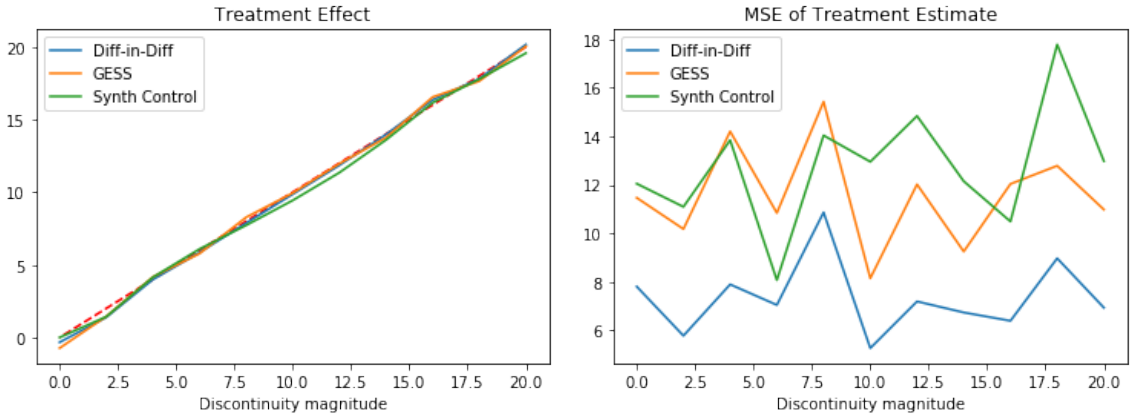


Fig. 6.3 Control identification methods applied to synthetic data assuming the true RDiT has been correctly identified. The left plot shows $\hat{\tau}$ at different τ magnitudes. The dashed red line indicates the true τ. The right plot shows the mean MSE of the individual treatment effect estimates.

While the results in Figure 6.3 assume that the true RDiT has been correctly identified, we now consider misidentification of the RDiT. In particular we vary the precision between s_τ and the true s from 0.3 to 1.0. Throughout all experiments we maintain a constant $\tau = 10$. The results in Figure 6.4 illustrate how misidentification of the RDiT results in incorrectly estimated $\hat{\tau}$.

Until now all the synthetic data has assumed that y are independent, conditional on x. Yet real world data often contains subsets of y that have correlated noise. We model this complexity by inducing a different data generating process in a predefined subset of the data, s_g. In particular, extending Equation 6.26, we generate,

$$y_i = x_i\gamma_y + \theta_i\tau + \varepsilon_i^{(y)} I(x_i \in s_g)t + u_i \tag{6.27}$$

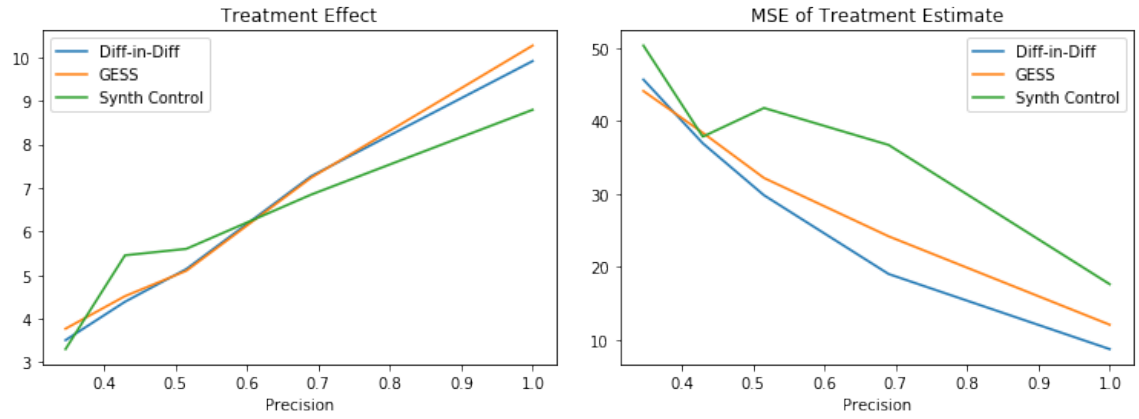


Fig. 6.4 Control identification methods applied to synthetic data assuming misidentification of the RDiT. The left plot shows $\hat{\tau}$ at varying levels of precision of s_τ. The right plot shows the mean MSE of the individual treatment effect estimates.

In our experiments we define s_g to be the union of the treated subset and a random subset of the non-treated records,

$$s_g = s_\tau \cup s_c,\ s_c \in \{D \setminus s_\tau\} \tag{6.28}$$

This intuitively reflects a real world setting where s_τ is differentially correlated with a subset of the remaining data.

Using this data generating process we rerun the experiments from Figure 6.3. Results are shown in Figure 6.5. Under these conditions GESS performs substantially better than either the standard DD control or the synthetic control. GESS's searching procedure identifies the superset of data corresponding to the altered generative data process. Both the standard DD and synthetic control procedures substantially overestimate τ because they assume a single data generative process and cannot properly disaggregate s_g from the remaining data.

Figure 6.5 also depicts results from the full-dimensional version of GESS. While both GESS algorithms are effective at identifying the correct subset, the full-dimensional version is more accurate in this case. This represents a bias-variance tradeoff between the two GESS approaches discussed in Section 6.3.2.

6.4.3 The California Smoking Legislation Study

In November 1988, California passed Proposition 99 anti-tobacco legislation which was intended to reduce tobacco and cigarette consumption within the state [73]. We consider the effect of this legislation by analyzing annual state-level data collected by [2]. This dataset includes smoking consumption per capita from 1970 to 2000 in 30 US states, including

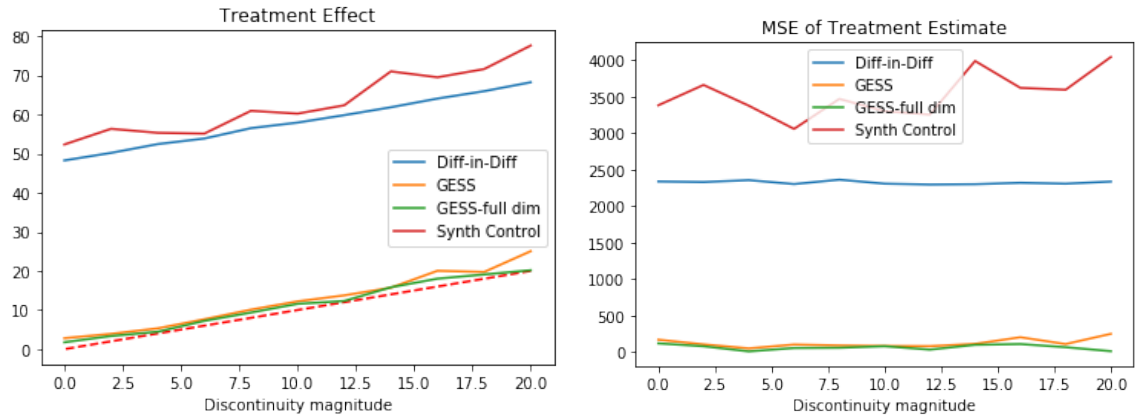


Fig. 6.5 Control identification methods applied to synthetic data assuming the true RDiT has been correctly identified. The data includes an alternative data generating process over s_g. The left plot shows $\hat{\tau}$ at different τ magnitudes. The dashed red line indicates the true τ. The right plot shows the mean MSE of the individual treatment effect estimates.

California. Figure 6.6 depicts smoking consumption per capita between 1970 and 2000 in California and three other states. Covariates in the dataset include per capita state personal income, average retail price of cigarettes, percentage of the population age 15-24, per capita beer consumption, and lagged variables of total smoking consumption in 1975, 1980, and 1988. Each covariate is quantized into four bins for our analysis in order to search over a discrete space. We construct a binary treatment variable which equals 1 only for California after Proposition 99 went into effect in January 1989.

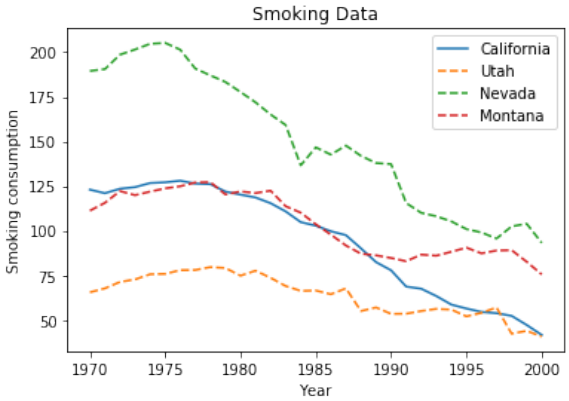


Fig. 6.6 Smoking consumption per capita between 1970 and 2000 in California and three other states

Method	$\hat{\tau}$	Significance
Diff-in-Diff	-10.94	Yes at 5%
Synthetic Control	-8.96	Yes at 5%
GESS	-6.67	Yes at 5%

Table 6.1 Estimated treatment effects in California smoking data using three control identification methods.

Previous literature has examined this dataset using both synthetic controls and DD where the treatment discontinuity is assumed to be in California at January 1989 [2]. We attempt to rediscover this discontinuity with SuDDDS. We apply the technique with each of the three *LLR* searching methods. All three methods perfectly identify the RDiT used in previous studies. Conditioning on the resulting s_τ we applied the three control subset identification methods. These all provided statistically significant estimates of $\hat{\tau}$ (shown in Table 6.1) all of which have the same direction and similar magnitudes.

While SuDDDS correctly identifies the RDiT when given binary treatment labels, we induce an increasingly difficult search problem by injecting synthetic noise into the treatment variable and creating non-binary treatment labels. We use the same approach from Section 5.4.5 for injecting the synthetic noise based on ρ.

Figure 6.7 depicts the precision, recall, and F-score of the discovered s_τ for each *LLR* search technique. While recall is relatively stable over increasing noise injection, the precision

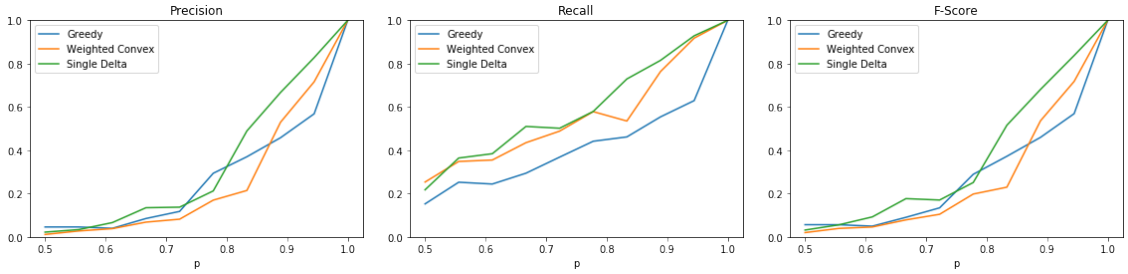


Fig. 6.7 Precision, recall, and F-score of RDiTs identified by SuDDDS in the California smoking data at varying amounts of injected noise. Three methods for optimizing *LLR* are compared.

reduces more rapidly. This is due to the small size of the true treatment subset relative to the entire dataset. Thus, even small errors in the estimated s_τ may lead to large degradation of the precision.

6.4.4 Traffic stop data

The Stanford Open Policing Project collects data on traffic stops by law enforcement agencies across the United States [143]. This data provides a unique perspective on the interactions between police and the public. By analyzing how that relationship changes we can gain insights into the effects of exogenous events (e.g. the passage of new legislation) on policing. This can help inform future policy as well as provide insight to guide police behavior.

We analyze data on the legal basis provided for each search by the police. All individuals in the United States are protected against illegal search and seizures by the Fourth Amendment of the Constitution, meaning that police must provide a legal justification for every search conducted as part of a traffic stop. In particular, there are two major reasons for searches provided in our data: consent and probable cause. Consent searches are those where an individual voluntarily allows the police to search their person or their property, without requiring the police or a court to compel such a search. Probable Cause searches are those where the police have a reasonable basis to believe that a crime is being, or has been committed. This justification is based on the police's judgment and permits a search that does not require consent from the individual [90].

The data were released as individual records of traffic stops with associated metadata and outcome data for each law enforcement agency. We consider state patrol law enforcement agencies who report the consent and probable cause search bases for traffic searches and who have nearly complete data between 2011-2016. We find nine states with the requisite data: Arizona (AZ), California (CA), Colorado (CO), Florida (FL), Massachusetts (MA), North Carolina (NC), Texas (TX), Wisconsin (WI), and Vermont (VT). We aggregate the individual record data to count-level data at a quarterly basis for each state. Figure 6.8 depicts quarterly count data of search bases for each state under investigation from 2011-2016. The data is heavily seasonal and varies substantially across states (see Figure 6.8) so we employ a base function, $f(x,t)$, with fixed effects over states as well as interaction between time and each state. We then run SuDDDS searching over states and search bases for the specified time period.

SuDDDS discovers a treated subset consisting of probable cause traffic searches in a single state, Colorado, after Q4 of 2012. We depict this DD by a dashed red line on the Colorado subplot in Figure 6.8. This discovered DD corresponds to Colorado's legalization of marijuana. Legalization occurs by a ballot initiative to amend the state constitution, which was passed in November 2012. In the context of traffic stops and searches, motorists were suddenly permitted to transport marijuana within the state. The sharp reduction in probable cause searches may have been caused by this policy change. Law enforcement officials who had probable cause to believe that a person was had marijuana in their vehicle would

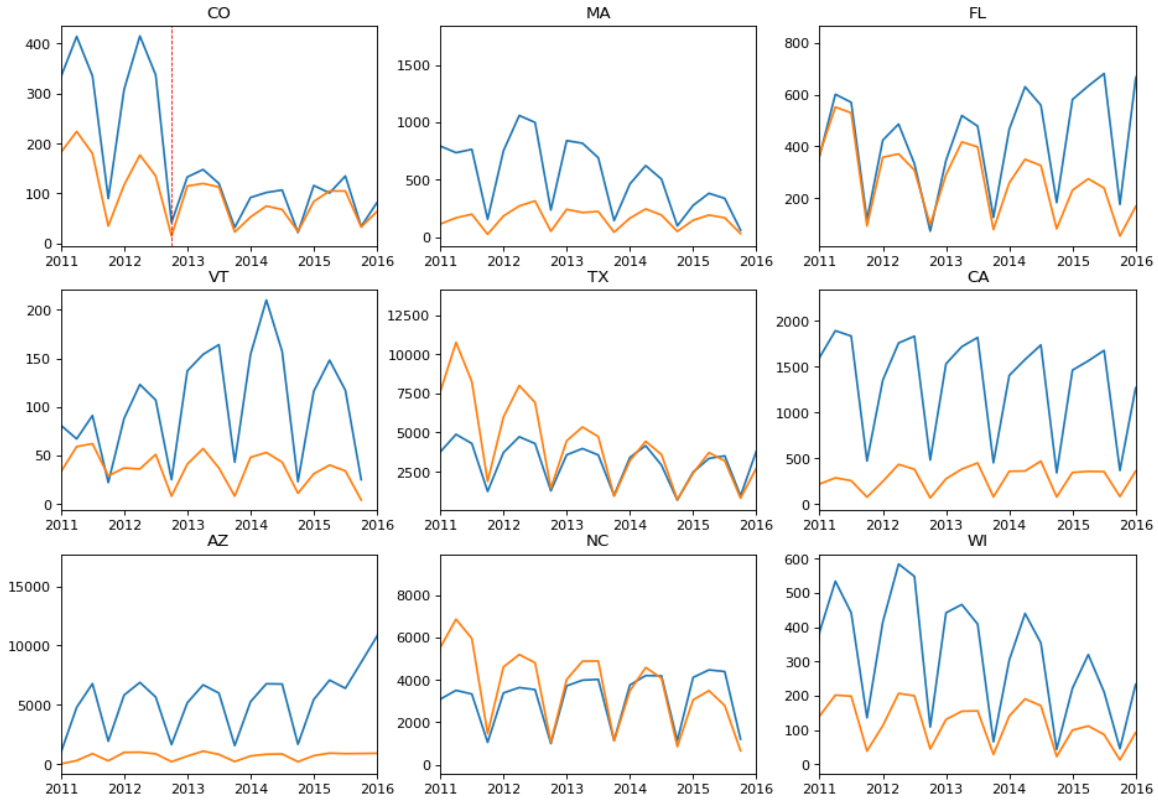


Fig. 6.8 Quarterly count data of search bases for each state under investigation from 2011-2016. The orange line indicates consent searches. The blue line indicates probable cause searches. The vertical dashed red line indicates the time of the DD discovered by SuDDDS.

previously have executed a probable cause search. But since transporting marijuana is no longer illegal after 2013, these searches ceased to occur.

Given this identified change, we investigate how the reduction in probable cause searches effected the racial composition of traffic stops and searches in Colorado. One important argument for marijuana legalization has been the disproportionate effect that marijuana criminalization has on racial minorities, particularly African Americans [119, 53]. We consider how the decrease in probable cause searches in Colorado has affected the distribution of police traffic searches across racial groups. In particular, we analyze the proportion of blacks and Hispanics searched as a percentage of the total number of individuals searched in each state. The dataset is equivalent to the data described above, consisting of quarterly data in the same nine states from 2011-2016. Figure 6.9 depicts aggregated data for this period and indicates the detected DD subset.

For both racial group we use all three of control identification methods to identify appropriate control subsets and compute treatment effects. We repeat this analysis using data

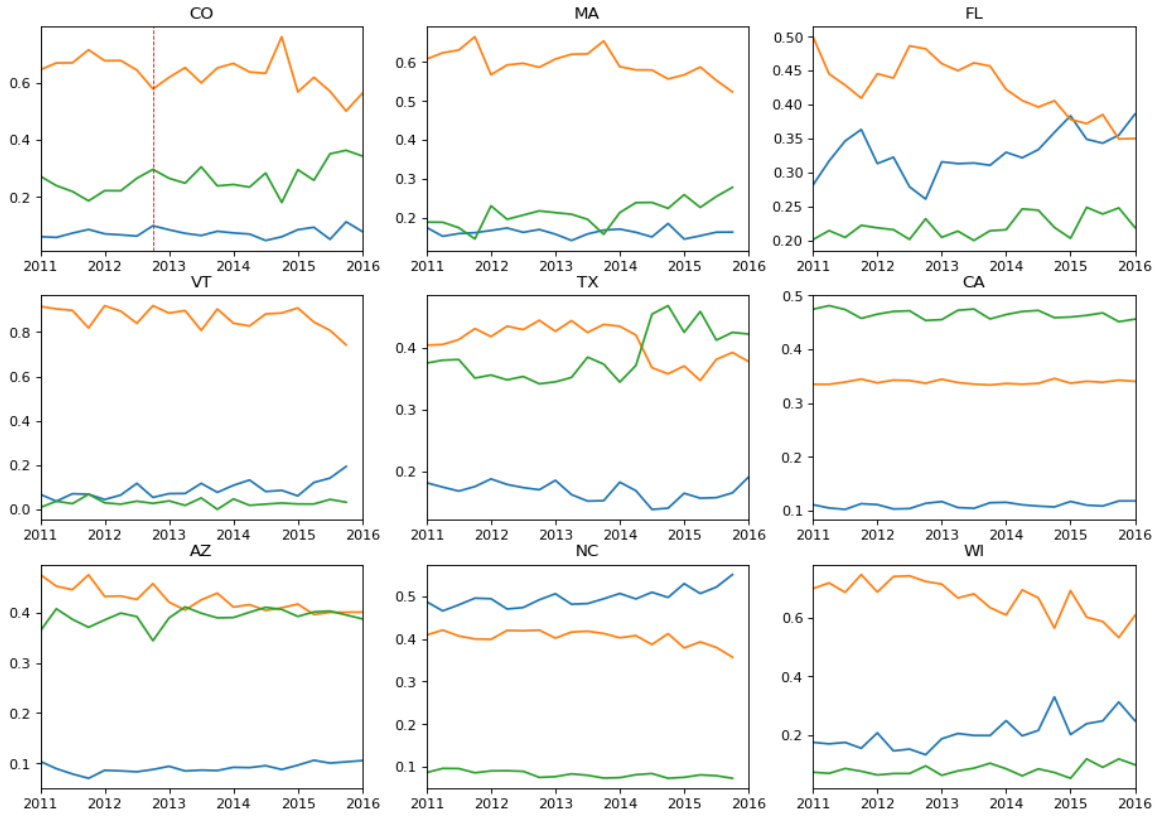


Fig. 6.9 Quarterly data of the proportion of whites, blacks, and Hispanics searched as a percentage of the total number of individuals searched in each state bases for each state under investigation from 2011-2016. The orange line indicates whites, the blue line indicates blacks, and the green line indicates Hispanics. The vertical dashed red line indicates the DD discovered by SuDDDS.

on the number of traffic stops by racial group and data on the number of traffic searches by racial group. Results of these analyses are shown in Tables 6.2 and 6.3.

All the estimated $\hat{\tau}$ values were close to zero, and randomization testing (described in Section 6.3.2) shows that none of the results were statistically significant. Since SuDDDS identified an RDiT these control methods provide unbiased estimates that the Colorado marijuana law did not impact the racial distribution of traffic searches. This null result reflects similar work on this dataset [143]. Additionally, it reflects other work on the legalization of marijuana in Colorado which suggests that the racial equity arguments advanced before the law may not been borne our in practice [44].

Method	Stopping		Searching	
	$\hat{\tau}$	Significance	$\hat{\tau}$	Significance
Diff-in-Diff	-0.0008	No	-0.0066	No
Synthetic Control	-0.0007	No	-0.0051	No
GESS	-0.0011	No	-0.0124	No

Table 6.2 Estimated treatment effects of the proportion of blacks stopped and searched by police in the traffic data. Results from all three control identification methods are compared.

Method	Stopping		Searching	
	$\hat{\tau}$	Significance	$\hat{\tau}$	Significance
Diff-in-Diff	-0.0042	No	-0.0053	No
Synthetic Control	-0.0032	No	-0.0027	No
GESS	-0.0002	No	-0.0057	No

Table 6.3 Estimated treatment effects of the proportion of Hispanics stopped and searched by police in the traffic data. Results from all three control identification methods are compared.

Chapter 7

Conclusions

In this thesis we examined a rich variety of machine learning techniques for identifying, characterizing, and exploiting changes in data. If you are looking for a summary of each chapter please see the introduction in Chapter 1. In concluding this work we will discuss some of the overarching themes woven throughout the preceding chapters, but not explicitly discussed.

Detection for Causal Inference Machine learning provides powerful models for detecting subtle patterns in data. For certain applications, such as those discussed in Chapter 4, mere detection of a particular pattern is an end in itself. Yet there is a substantial body of work using detection techniques as the first stage of a larger process, such as "detection for prediction" where automatically discovered patterns are used as features for prediction models (e.g. Jean et al. [78] among others).

Work in this thesis considers a conceptual framework we might call "detection for causal inference," whereby we use automatically discovered changes in data for as the building blocks of a subsequent causal model. For example, in Chapter 2 we use change surfaces as the basis for counterfactual prediction. Similarly, in Chapters 5 and 6 we employ anomalous pattern detection to identify natural experiments that allow us to compute treatment effects. Data changes provide a natural opening for causal inference since they signal shifts in the data distribution, allowing us a window – ever so briefly – into possible counterfactual worlds.

Machine learning and econometrics As noted in Chapter 3 there has recently been a surge of interest at the intersection of machine learning and econometrics. The chapters in this thesis on causal inference all contribute to this literature. In particular the models we present in Chapters 5 and 6 pioneer novel ways of thinking about the connection between machine learning and econometrics by employing automated search techniques for identifying natural

experiments. The framework we develop in those chapters is powerful but nascent, leaving much room for further development in this domain.

Real real-world data This thesis is submitted in fulfillment of a PhD in both machine learning and public policy. Throughout these chapters we emphasize the use of data resulting from real processes existing in the world, specifically those relating to public policy and public health. These data are complex and messy. They do not obviously conform to the data generating assumptions of any machine learning algorithm. Yet it is precisely this mismatch that makes such data important. The statistical algorithms in this thesis were created to serve public policy experts and must be usefully applied to the data those practitioners deal with on a day-to-day basis.

We exclusively use data available to the general public. While proprietary data can provide important insights to policy questions, research resulting from those data are difficult (if not impossible) to replicate. This hinders the ability of researchers to validate and expand on past results. Additionally, access to proprietary data is unequally distributed and its use in public science tends to favor well-financed and established institutions. For these reasons all data in this thesis can be accessed online, as of the publication of the thesis. References to online sources are provided in the relevant chapters.

References

[1] Abadie, A., Angrist, J., and Imbens, G. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, 70(1):91–117.

[2] Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505.

[3] Abadie, A., Diamond, A., and Hainmueller, J. (2011). Synth: An r package for synthetic control methods in comparative case studies. *Journal of Statistical Software*, 42(13).

[4] Abadie, A., Diamond, A., and Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510.

[5] AccuWeather (2013). Photos: Tornadoes ransacked Brooklyn, Queens in 2010.

[6] Adams, R. P. and MacKay, D. J. (2007). Bayesian online changepoint detection. *stat*, 1050:19.

[7] Agarwal, D., McGregor, A., Phillips, J. M., Venkatasubramanian, S., and Zhu, Z. (2006). Spatial scan statistics: approximations and performance study. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 24–33.

[8] Aminikhanghahi, S. and Cook, D. J. (2017). A survey of methods for time series change point detection. *Knowledge and information systems*, 51(2):339–367.

[9] Amjad, M., Shah, D., and Shen, D. (2018). Robust synthetic control. *The Journal of Machine Learning Research*, 19(1):802–852.

[10] Anderson, M., Dobkin, C., and Gross, T. (2012). The effect of health insurance coverage on the use of medical services. *AEJ: Economic Policy*, 4(1).

[11] Anderson, M. L., Dobkin, C., and Gross, T. (2014). The effect of health insurance on emergency department visits: Evidence from an age-based eligibility threshold. *Review of Economics and Statistics*, 96(1):189–195.

[12] Angrist, J. D. and Krueger, A. B. (1999). Empirical strategies in labor economics. In *Handbook of labor economics*, volume 3, pages 1277–1366. Elsevier.

[13] Angrist, J. D. and Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

[14] Ashenfelter, O. and Card, D. (1984). Using the longitudinal structure of earnings to estimate the effect of training programs. Technical report, National Bureau of Economic Research.

[15] Athey, S. (2015). Machine learning and causal inference for policy evaluation. In *Proceedings of the 21th ACM SIGKDD Conference*, pages 5–6. ACM.

[16] Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proc. of the National Academy of Sciences*, 113(27):7353–7360.

[17] Athey, S. and Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497.

[18] Aue, A. and Horváth, L. (2013). Structural breaks in time series. *Journal of Time Series Analysis*, 34(1):1–16.

[19] Bertanha, M. (2016). Regression discontinuity design with many thresholds.

[20] Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, 119(1):249–275.

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

[22] Bilinski, A. and Hatfield, L. A. (2018). Seeking evidence of absence: reconsidering tests of model assumptions. *arXiv preprint arXiv:1805.03273*.

[23] Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., and Scott, S. L. (2015). Inferring causal impact using bayesian structural time-series models. *Annals of Applied Statistics*, 9:247–274.

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

[25] Card, D., Mas, A., and Rothstein, J. (2008). Tipping and the dynamics of segregation. *The Quarterly Journal of Economics*, 123(1):177–218.

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

[27] CBS New York (2016a). Discoloration in water rankles upper Manhattan residents.

[28] CBS New York (2016b). Severe storms follow extreme heat for tri-state area.

[29] Census Bureau, U. S. (1999). United states historical census data. https://www.census.gov/hhes/www/income/data/historical/state/. Accessed: 2016-4-10.

[30] Census Bureau, U. S. (2014a). American community survey 1-year estimates. http://factfinder.census.gov/. Accessed: 2016-4-10.

[31] Census Bureau, U. S. (2014b). American community survey 5-year estimates. http://factfinder.census.gov/. Accessed: 2016-4-10.

[32] Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):15.

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

[34] Cheng, K.-W., Chen, Y.-T., and Fang, W.-H. (2015). Video anomaly detection and localization using hierarchical feature representation and Gaussian process regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2909–2917.

[35] 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.

[36] City, N. Y. (2016). Water lead test kit request. http://www1.nyc.gov/nyc-resources/service/1266/water-lead-test-kit-request. Accessed: 2016-4-10.

[37] City of New York (2017). New York City open data.

[38] City of New York Office of the Mayor (2017). HealingNYC: Preventing overdoses, saving lives.

[39] Control, C. F. D. and Prevention (2016). Epidemiology and prevention of vaccine-preventable diseases.

[40] Dalziel, B. D., Bjørnstad, O. N., van Panhuis, W. G., Burke, D. S., Metcalf, C. J. E., and Grenfell, B. T. (2016). Persistent chaos of measles epidemics in the prevaccination united states caused by a small change in seasonal transmission patterns. *PLoS Comput Biol*, 12(2):e1004655.

[41] Del Pia, A., Dey, S. S., and Molinaro, M. (2014). Mixed-integer quadratic programming is in NP. *Mathematical Programming*, pages 1–16.

[42] Dimmery, D. (2016). *rdd: Regression Discontinuity Estimation*. R package v0.57.

[43] Doudchenko, N. and Imbens, G. W. (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Technical report, National Bureau of Economic Research.

[44] Drug Policy Alliance (2018). From prohibition to progress: A status report on marijuana legalization.

[45] Duczmal, L., Cancado, A. L., Takahashi, R. H., and Bessegato, L. F. (2007). A genetic algorithm for irregularly shaped spatial scan statistics. *Computational Statistics and Data Analysis*, 52:43–52.

[46] 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(5453):667–670.

[47] Editorial Board, T. (2016). Poisoned water in newark schools. *New York Times*.

[48] Ferman, B., Pinto, C., and Possebom, V. (2017). Cherry picking with synthetic controls.

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

[50] 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*.

[51] 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.

[52] Gay, M. (2016). Elevated levels of lead found in water of some vacant public-housing apartments. *Wall Street Journal*.

[53] Governor of New York State (2019). Governor cuomo signs legislation decriminalizing marijuana use.

[54] Guan, Z. (2004). A semiparametric changepoint model. *Biometrika*, pages 849–862.

[55] Hahn, J., Todd, P., and Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1).

[56] Hansen, B. E. (2000). Sample splitting and threshold estimation. *Econometrica*, 68(3):575–603.

[57] Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2016). Counterfactual prediction with deep instrumental variables networks. *arXiv preprint arXiv:1612.09596*.

[58] Hausman, C. and Rapson, D. S. (2018). Regression discontinuity in time: Considerations for empirical applications. *Annual Review of Resource Economics*, 10:533–552.

[59] Hawkins, D. M., Qiu, P., and Kang, C. W. (2003). The changepoint model for statistical process control. *Journal of quality technology*, 35(4).

[60] Hawkins, D. M. and Zamba, K. (2005). A change-point model for a shift in variance. *Journal of Quality Technology*, 37(1):21.

[61] Herlands, W., McFowland, E., Wilson, A. G., and Neill, D. (2018a). Gaussian process subset scanning for anomalous pattern detection in non-iid data. In *21st International Conference on Artificial Intelligence and Statistics, AISTATS 2018*.

[62] Herlands, W., McFowland III, E., Wilson, A. G., and Neill, D. B. (2018b). Automated local regression discontinuity design discovery. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1512–1520. ACM.

[63] Herlands, W., Neill, D. B., Nickisch, H., and Wilson, A. G. (2019). Change surfaces for expressive multidimensional changepoints and counterfactual prediction. *Journal of Machine Learning Research*, 20(99):1–51.

[64] Herlands, W., Wilson, A., Nickisch, H., Flaxman, S., Neill, D., Van Panhuis, W., and Xing, E. (2016). Scalable gaussian processes for characterizing multidimensional change surfaces. In *Artificial Intelligence and Statistics*, pages 1013–1021.

[65] Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

[66] Hodge, V. and Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126.

[67] Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

[68] Horváth, L. and Rice, G. (2014). Extensions of some classical methods in change point analysis. *Test*, 23(2):219–255.

[69] Horwitz, L. I., Green, J., and Bradley, E. H. (2010). Us emergency department performance on wait time and length of visit. *Annals of emergency medicine*, 55(2):133–141.

[70] Hsu, Y.-C., Shen, S., et al. (2016). Testing for treatment effect heterogeneity in regression discontinuity design. Technical report, Academia Sinica, Taiwan.

[71] III, E. M., Somanchi, S., and Neill, D. B. (2018). Efficient discovery of heterogeneous treatment effects in randomized experiments via anomalous pattern detection. *Working paper*.

[72] Imbens, G. W. and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of econometrics*, 142(2):615–635.

[73] Institute of Governmental Studies, University of California, Berkeley (2008). Proposition 99.

[74] Ivanoff, B. G. and Merzbach, E. (2010). Optimal detection of a change-set in a spatial poisson process. *The Annals of Applied Probability*, pages 640–659.

[75] Jacob, R., Zhu, P., Somers, M.-A., and Bloom, H. (2012). A practical guide to regression discontinuity. *MDRC*.

[76] 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*.

[77] Jarrett, R. (1979). A note on the intervals between coal-mining disasters. *Biometrika*, pages 191–193.

[78] Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794.

[79] Johansson, F., Shalit, U., and Sontag, D. (2016). Learning representations for counterfactual inference. In *ICML*, pages 3020–3029.

[80] Kapur, P. K., Pham, H., Gupta, A., and Jha, P. C. (2011). *Change-Point Models*, pages 171–213. Springer London, London.

[81] Kaul, A., Klößner, S., Pfeifer, G., and Schieler, M. (2015). Synthetic control methods: Never use all pre-intervention outcomes together with covariates.

[82] Keshavarz, H., Scott, C., and Nguyen, X. (2018). Optimal change point detection in gaussian processes. *Journal of Statistical Planning and Inference*, 193:151–178.

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

[84] Kowalska, K. and Peel, L. (2012). Maritime anomaly detection using Gaussian process active learning. In *Information Fusion (FUSION), 2012 15th International Conference on*, pages 1164–1171. IEEE.

[85] Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics-Theory and methods*, 26(6):1481–1496.

[86] Kulldorff, M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 164(1):61–72.

[87] Kulldorff, M., Huang, L., Pickle, L., and Duczmal, L. (2006). An elliptic spatial scan statistic. *Statistics in Medicine*, 25(22):3929–3943.

[88] 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.

[89] Lee, D. S. and Lemieuxa, T. (2010). Regression discontinuity designs in economics. *Journal of economic literature*, 48(2):281–355.

[90] Lemons, B. R. (2002). Searching a vehicle without a warrant.

[91] Lindo, J. M., Sanders, N. J., and Oreopoulos, P. (2010). Ability, gender, and performance standards: Evidence from academic probation. *AEJ: Applied Economics*, 2(2).

[92] 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*.

[93] MacKay, D. J. (1998). Introduction to gaussian processes. *NATO ASI Series F Computer and Systems Sciences*, 168:133–166.

[94] MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

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

[96] 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.

[97] McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of econometrics*, 142(2).

[98] Minka, T. P. (2001). Automatic choice of dimensionality for pca. In *Advances in neural information processing systems*, pages 598–604.

[99] Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998). Log gaussian cox processes. *Scandinavian journal of statistics*, 25(3):451–482.

[100] Mullainathan, S. and Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

[101] Neill, D. B. (2009). Expectation-based scan statistics for monitoring spatial time series data. *International Journal of Forecasting*, 25:498–517.

[102] Neill, D. B. (2012). Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):337–360.

[103] Neill, D. B., McFowland III, E., and Zheng, H. (2013). Fast subset scan for multivariate event detection. *Statistics in medicine*, 32(13):2185–2208.

[104] Neill, D. B. and Moore, A. W. (2004). Rapid detection of significant spatial clusters. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 256–265. ACM.

[105] Neill, D. B., Moore, A. W., Sabhnani, M. R., and Daniel, K. (2005). Detection of emerging space-time clusters. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 218–227.

[106] 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*.

[107] NYC Department of Education (2017). Data about schools.

[108] Page, E. (1954). Continuous inspection schemes. *Biometrika*, 41(1/2):100–115.

[109] Pichardo, C. (2016). Brown water pours from faucets uptown after maintenance work, DEP says.

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

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

[112] Rambachan, A. and Roth, J. (Working Paper). An honest approach to parallel trends (job market paper).

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

[114] Rasmussen, C. E. and Ghahramani, Z. (2001). Occam's razor. In *Advances in neural information processing systems*, pages 294–300.

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

[116] Rasmussen, C. E. and Nickisch, H. (2016). GPML Matlab code version 4.0.

[117] Reardon, S. F. and Robinson, J. P. (2012). Regression discontinuity designs with multiple rating-score variables. *JREE*, 5(1):83–104.

[118] Reece, S., Garnett, R., Osborne, M., and Roberts, S. (2015). Anomaly detection and removal using non-stationary gaussian processes. *arXiv preprint arXiv:1507.00566*.

[119] Resing, C. (2019). Marijuana legalization is a racial justice issue.

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

[121] Ross, G. J. and Adams, N. M. (2012). Two nonparametric control charts for detecting arbitrary distribution changes. *Journal of Quality Technology*.

[122] Ross, G. J., Tasoulis, D. K., and Adams, N. M. (2011). Nonparametric monitoring of data streams for changes in location and scale. *Techno.*, 53(4).

[123] Rousseeuw, P. J. and Leroy, A. M. (2005). *Robust regression and outlier detection*, volume 589. John Wiley & Sons.

[124] Rousseeuw, P. J. and Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411):633–639.

[125] Rubin, D. B. (1986). Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396):961–962.

[126] Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

[127] Saatçi, Y. (2011). *Scalable inference for structured Gaussian process models*. PhD thesis, University of Cambridge.

[128] 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.

[129] Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471.

[130] Schölkopf, B. and Smola, A. J. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

[131] Schulam, P. and Saria, S. (2017). Reliable decision support using counterfactual models. In *Advances in Neural Information Processing Systems*, pages 1697–1708.

[132] Scott, A. J. and Knott, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, pages 507–512.

[133] Scottish Mining (1887). Coal mines regulation act.

[134] Sharkey, P. and Killick, R. (2014). Nonparametric methods for online changepoint detection. Technical Report STOR601, Lancaster University.

[135] Sharma, A. (2016). Necessary and probably sufficient test for finding valid instrumental variables. Technical report, working paper, Microsoft Research, NY.

[136] Sharma, A., Hofman, J. M., and Watts, D. J. (2016). Split-door criterion for causal identification: Automatic search for natural experiments. *arXiv:1611.09414*.

[137] Shirota, S. and Gelfand, A. E. (2016). Inference for log gaussian cox processes using an approximate marginal posterior. *arXiv preprint arXiv:1611.10359*.

[138] Smith, M., Reece, S., Roberts, S., Psorakis, I., and Rezek, I. (2014). Maritime abnormality detection using gaussian processes. *Knowledge and information systems*, 38(3):717–741.

[139] Speakman, S., Somanchi, S., McFowland III, E., and Neill, D. B. (2016). Penalized fast subset scanning. *Journal of Computational and Graphical Statistics*, 25(2):382–404.

[140] Stegle, O., Fallert, S. V., MacKay, D. J., and Brage, S. (2008). Gaussian process robust regression for noisy heart rate data. *IEEE Transactions on Biomedical Engineering*, 55(9):2143–2151.

[141] Tartakovsky, A. G., Polunchenko, A. S., and Sokolov, G. (2013). Efficient computer network anomaly detection by changepoint detection methods. *IEEE Journal of Selected Topics in Signal Processing*, 7(1):4–11.

[142] Trzeciak, S. and Rivers, E. (2003). Emergency department overcrowding in the united states: an emerging threat to patient safety and public health. *Emergency medicine journal*, 20(5):402–405.

[143] University, S. (2016). The stanford open policing project.

[144] US CDC (2017). WONDER database.

[145] US Department of Health and Human Services (2016). The opioid epidemic: By the numbers.

[146] Van der Klaauw, W. (2008). Regression–discontinuity analysis: a survey of recent developments in economics. *Labour*, 22(2):219–245.

[147] 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.

[148] Verbitsky-Savitz, N. and Raudenbush, S. W. (2012). Causal inference under interference in spatial settings: a case study evaluating community policing program in chicago. *Epidemiological Methods*, 1:106–130.

[149] Water Study, F. (2015). Flint water study: Articles in the press. http://flintwaterstudy.org/articles-in-the-press/. Accessed: 2016-4-10.

[150] 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.

[151] 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.

[152] 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.

[153] 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.

[154] Wilson, A. and Nickisch, H. (2015). Kernel interpolation for scalable structured gaussian processes (kiss-gp). In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1775–1784.

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

[156] Wilson, A. G., Hu, Z., Salakhutdinov, R., and Xing, E. P. (2016a). Deep kernel learning. In *Artificial Intelligence and Statistics*, pages 370–378.

[157] Wilson, A. G., Hu, Z., Salakhutdinov, R. R., and Xing, E. P. (2016b). Stochastic variational deep kernel learning. In *Advances in Neural Information Processing Systems*, pages 2586–2594.

[158] Wong, V. C., Steiner, P. M., and Cook, T. D. (2013). Analyzing regression-discontinuity designs with multiple assignment variables: A comparative study of four estimation methods. *Journal of Educational and Behavioral Statistics*.

[159] Wu, M., Song, X., Jermaine, C., Ranka, S., and Gums, J. (2009). A LRT framework for fast spatial anomaly detection. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 887–896.

[160] Yu, S., Tranchevent, L.-C., De Moor, B., and Moreau, Y. (2013). *Kernel-based data fusion for machine learning*. Springer.

[161] Zhang, Z. and Neill, D. B. (2016). Identifying significant predictive bias in classifiers. *arXiv preprint arXiv:1611.08292*.