DATA-DRIVEN DISCOVERY IN POLITICAL SCIENCE

by

Marc Thomas Ratkovic

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Political Science)

at the

UNIVERSITY OF WISCONSIN–MADISON

2011

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Data-Driven Discovery in Political Science

submitted to the Graduate School of the
University of Wisconsin-Madison
in partial fulfillment of the requirements for the
degree of Doctor of Philosophy

By

Marc Ratkovic

Date of final oral examination: July 25, 2011

Month and year degree to be awarded: August 2011

The dissertation is approved by the following members of the Final Oral Committee:

David Weimer, Professor, Political Science and Public Affairs
Charles Franklin, Professor, Political Science
Alexander Tahk, Assistant Professor, Political Science
Erik Nordheim, Professor, Statistics
Kosuke Imai, Professor, Politics, Princeton University

To my grandfather, Wilfred Elmer Schmaltz.

Where he rests, the Cubs win every World Series.

ACKNOWLEDGMENTS

While the title page lists me as author, this dissertation is truly the product of the support, encouragement, assistance, and love of the many wonderful people life has blessed me with. First, I would like to acknowledge my grandparents, Wilfred and Josephine Schmaltz, and my parents, Thomas and Laura Ratkovic, for their unwavering support and patience; my adivsor, David Weimer, for guidance and patience; and my supervisor, Kosuke Imai, for opportunity and guidance. This project would not have materialized without their support.

The environments at the University of Wisconsin-Madison and Princeton have provided wonderful arenas for growth and feedback. I would particularly like to thank the Models and Data Group at the University of Wisconsin and the Political Methodology Group at Princeton, for sharpening many of the ideas within. Charles Franklin and Erik Nordheim have provided guidance throughout this project, and Alex Tahk has helped greatly in the final stages. I thank them all gratefully.

My many friends have helped this dissertation develop. The space is too short to list all of them, but I would like to thank in particular: my best friends, my brothers Paul and Peter, for encouragement and emotional support throughout the entire process; John and Esther Fitzgerald for ongoing friendship; Walter Mebane, Ilia Murtazashvili, and Stephane Lavertu for unwavering mentorship and introducing me to the discipline; Kevin Eng for the many wonderful, insightful conversations about statistics, observational versus experimental inference, episetmology, and Borges; Jess Clayton, for being an anchor; Mehreen Zalia-Malik for the poetry; Jim Jazwiecki, Drew Stathus, and Doug Rosso for computer help mixed with great friendship; and Jason Ardanowski, Kyle Marquardt, Dimitri Kelly, Jake Neiheisel, and John Strand, for the opportunity to help introduce newer scholars to the field.

Heartfelt thanks go to Holly Bedwell, Jenn Lindsay, and Mimi King, the women who have joined my life through different parts of this project.

The ideas in this dissertation are the result of many fascinating conversations and presentations, and owes a debt to each. All mistakes are my own.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Appendix
Figure Page

DATA-DRIVEN DISCOVERY IN POLITICAL SCIENCE

Marc Thomas Ratkovic

Under the supervision of Professor of Public Affairs and Political Science David Leo Weimer
At the University of Wisconsin-Madison

Commonly used approaches in political methodology do not suffice when the number of variables grows large. Rather than use data to test hypotheses, the first chapter presents a series of methods that allow the researcher to identify the relevant predictive variables within the data. The methods allow for fitting models consisting of hundreds of variables, while selecting only a small subset. Interaction effects, normally left unmodeled, and small effects can be identified. This is done by recasting the problem as one of variable selection. Smoothing splines are incorporated to allow for nonlinearities in the data.

The second chapter uses variable selection methods to identify treatment effect heterogeneity, by placing separate sparsity constraints over differing causal heterogeneity parameters of interest within a support vector machine classifier. As confirmed in simulation studies, the proposed method tends to yield lower false discovery rate than commonly used alternatives. For empirical illustrations, I apply the proposed method to randomized field experiments from political science and economics.

The third chapter develops a sequential segmentation spline method that identifies the location and number of changepoints in a series of observations with a smooth time component, using a modified BIC statistic as a stopping rule. I explore the method in a large-n, unbalanced panel setting with George W. Bush's approval data, a small-n time series with median DW-NOMINATE scores for each Congress over time, and a series of simulations.

The method performs favorably in terms of visual inspection, residual properties, and event detection relative to extant smoothers.

The final chapter identifies effects that correspond with jurisdictional boundaries, while allowing for smooth geographic correlation. The method combines smoothing splines with variable selection, identifying non-zero jurisdiction-specific effects. The proposed method offers researchers the ability to fit a large number of effects, from dozens to hundreds, while only returning the most relevant effects. Simulations show that the method has a low false discovery rate, and is quite powerful. Applications to African GDP data, US voting patterns, and US crime rates illustrate the proposed method's efficacy and use.

ABSTRACT

Commonly used approaches in political methodology do not suffice when the number of variables grows large. Rather than use data to test hypotheses, the first chapter presents a series of methods that allow the researcher to identify the relevant predictive variables within the data. The methods allow for fitting models consisting of hundreds of variables, while selecting only a small subset. Interaction effects, normally left unmodeled, and small effects can be identified. This is done by recasting the problem as one of variable selection. Smoothing splines are incorporated to allow for nonlinearities in the data.

The second chapter uses variable selection methods to identify treatment effect heterogeneity, by placing separate sparsity constraints over differing causal heterogeneity parameters of interest within a support vector machine classifier. As confirmed in simulation studies, the proposed method tends to yield lower false discovery rate than commonly used alternatives. For empirical illustrations, I apply the proposed method to randomized field experiments from political science and economics.

The third chapter develops a sequential segmentation spline method that identifies the location and number of changepoints in a series of observations with a smooth time component, using a modified BIC statistic as a stopping rule. I explore the method in a large-n, unbalanced panel setting with George W. Bush's approval data, a small-n time series with median DW-NOMINATE scores for each Congress over time, and a series of simulations.

The method performs favorably in terms of visual inspection, residual properties, and event detection relative to extant smoothers.

The final chapter identifies effects that correspond with jurisdictional boundaries, while allowing for smooth geographic correlation. The method combines smoothing splines with variable selection, identifying non-zero jurisdiction-specific effects. The proposed method offers researchers the ability to fit a large number of effects, from dozens to hundreds, while only returning the most relevant effects. Simulations show that the method has a low false discovery rate, and is quite powerful. Applications to African GDP data, US voting patterns, and US crime rates illustrate the proposed method's efficacy and use.

Chapter 1

An Introduction to Data-Driven Hypothesis Generation

1.1 Introduction

Uncovering causal relations in data is central to social science. Well-developed methods exist to test a small number of pre-specified hypotheses, while accounting for known confounders (e.g., King, 1998). As the complexity of data and the number of variables grow, estimating scores or hundreds of parameters grows infeasible. Infeasible, though, is not the same as uninteresting. Including these additional parameters can lead to a much more subtle depiction of the data than available through current methods. I propose a series of methods that allow the estimation of complex models while producing interpretable results.

Rather than conduct simultaneous tests of large numbers of variables, this dissertation concerns itself with a wholly separate approach to using data: that of discovery rather than inference. Common practice relies on theory-driven variable selection, where a priori theory generates some empirical implication, and inference is conducted. This dissertation turns this entire process on its head. Rather than assert a model and a hypothesis, which is then tested, I propose methods that allow for the selection of a sparse model. The sparse model is selected such that most coefficients are set to zero, and estimated so as to minimize a predictive criterion. The proposed methods select hypotheses that the researcher should have asked. Balancing model fit against model size guards against overfitting.

The proposed methods allow the fitting of complex models, while returning parsimonious results. A large number of variables can be considered, hundreds in many cases, with as few as ten selected. This dissertation presents just a few of the cases where these methods

may be helpful. A prominent field experiment conducted in New Haven in 1998 consisted of four crossed factors aimed at increasing voter turnout: one of three appeals (civic duty, neighborhood solidarity, or a close election), zero to three mailings sent, seven possible phone messages, and a personal visit (Gerber and Green, 2000). All possible interactions among these factors produce 279 different effects. Analyses that only considers main effects leaves these interactions unmodeled and unexplored. Instead of "fishing" for significant interactions, I consider all main and interactive effects simultaneously, selecting the most powerful effects.

The proposed methods differ from prominent work in political science that have fit complex models to data (Green and Kern, 2010a; Beck *et al.*, 2000; Hill and McCulloch, 2007). The most commonly used methods fit smooth, high-dimensional curves to data, as in with neural networks (Ripley, 1996) or sums-of-trees models (Breiman, 2001; Chipman *et al.*, 2010). These models are difficult to interpret; no coefficients are returned that can be relayed back to some independent variables of interest in a straightforward manner. The proposed methods remedy this by producing coefficients for a (possibly vast) number of variables. Setting most of these coefficients to zero serves to select relevant variables, and their coefficients can be interpreted in a normal manner. I refer to these methods, as applied by political and other social scientists, as "data-driven hypothesis generation," so as to separate it from the prediction problem. These selected variables can either be tested on a different dataset, or used to further theory by suggesting previously unrecognized stylized facts.

Two sets of effects are most amenable to this approach: small effects and interaction effects. Both are considered in this dissertation. Large effects are already well-known: previous voters are more likely to vote in the future; local geographic conditions and population density predict growth; previous Presidential approval predicts current approval; and so on. Smaller effects, in the presence of large effects, are difficult to distinguish from noise (Gelman and Weakliem, 2007). This is especially the case when the researcher has many plausible small effects over which she is agnostic. This dissertation proposes a method that explicitly

separates out small effects from known large effects, so the former can be identified and not washed out by the latter.

The second set of effects most open to this approach are interaction effects. Though main effects are the most important, in terms of explanation, interaction effects are often the most interesting, especially in identifying when the effect of a key covariate varies systematically across the values of another covariate. Within the context of experimental or quasi-experimental data, interactions can be used to parameterize causal heterogeneity. For example, a treatment may have an effect that varies systematically across different demographic subgroups. The proposed method uncovers that a job training program was least effective for black recipients with no degree, but most effective for older, unemployed recipients. This level of fine-grained analysis cannot be accommodated by existing methods. Interaction effects can also be used to capture effect heterogeneity, whereby an outcome can vary systematically from one jurisdiction to the next. The proposed methods allow for the consideration of all interaction effects up to any arbitrary level.

I accomplish this through recasting practical problems within a variable selection framework. Every researcher has had to address the question of which variables to include in a model. A host of questions are nearly always left unanswered: why no interactions, or only the small number provided? Why no quadratic or cubic terms? The variable selection literature strives to provide a rigorous, and data-driven, answer to this question. I focus primarily on variable selection through the Least Absolute Selection and Shrinkage Operator, or LASSO (Tibshirani, 1996).

The LASSO is a penalized regression method, where a linear model is fit subject to a constraint on the sum of the absolute values of the parameters. As discussed below, this has the desirable property of producing point estimates of *precisely* zero for most effects. In practice, the method and its extensions have been shown to be a powerful means to identify a meaningful subset of variables. The LASSO has generated a vast literature, across fields from statistics and computer science to biology and public policy (Hesterberg *et al.*, 2008). Political scientists have been remarkably silent in this field. I introduce political scientists

to many of these insights, showing their utility, as well as extending the methods broadly available.

While researchers often have some idea *which* variables to include, there may be concerns over unmodeled nonlinearities (Wahba, 1990; Beck and Jackman, 1997; Keele, 2008). The proposed methods allow for both smoothing and variable selection. By couching both sets of methods, variable selection and smoothing, within a regularization framework, the means to both model arbitrarily smooth curves while selecting from a set of variables becomes clear. Linearity assumptions and variable inclusion concerns can both be greatly ameliorated within this framework.

The dissertation consists of four chapters. This introductory chapter provides an overview and addresses some of the theoretical concerns of such an approach. In common practice, theory is used to generate a hypothesis of interest. An outcome of interest is assumed linear in this variable, as well as a set of known predictors. The model is fit and a p-value is used to test whether the null hypothesis of no effect can be rejected. Rather than assume a model, the proposed methods selects a sparse model from a large set of possible models. The means for doing so are developed in the first chapter.

The remainder of the dissertation consists of three papers. Each paper illustrates the utility of variable selection in addressing questions of concern to political scientists. The first paper considers problems of causal effects. The reporting of an average treatment effect ignores causal heterogeneity: the treatment effect may vary importantly and systematically across different subgroups. Considering causal heterogeneity as a variable selection problem allows these interactions to be identified in a manner both useful and informative (Imai and Strauss, 2011). When all subgroups are considered, the number of hypotheses may approach or even outgrow the sample size. The proposed method identifies these effects in a statistically rigorous manner, while considering many more hypotheses than current practice allows.

The next two papers deal with problems of data-driven discovery. The first paper in this section builds on the literature of identifying change points in time series (Calderia and Zorn,

1998; Spirling, 2007b). Rather than assume a linear model, I fit a model where the mean function is a smooth curve in time. An algorithm for uncovering these change points, and a modified BIC statistic that serves as a stopping rule, are both introduced. The proposed method is applied to President George W. Bush's approval data and to DW-NOMINATE ideology scores.

The final paper extends the change point problem to its two-dimensional variant. Though existing methods allow the testing of a single political boundary as a discontinuity (Keele and Titiunik, 2011), the proposed method identifies effects that respect state boundaries. A model additive in a smooth geographic component and a discontinuous state-specific effect is fit. "Red" and "blue" states in the 2008 United States presidential race are identified, as are high and low growth states in Africa, all while accounting for the smooth progression of socioeconomic outcomes across a geography.

The proposed methods allow political scientists to consider a much broader array of variables and interactions than under current practice. By fitting large models, but returning parsimonious results, the proposed methods can be used to identify effects even in the presence of hundreds or thousands of variables. Complex correlations with vote choice, coalition duration, and onset of war can all be examined in a manner both statistically rigorous and easy to interpret. The proposed methods make identifying normally unmodeled complexities and nonlinearities in our data tractable, in a statistically rigorous manner. Since most of the methods in this dissertation are new, or only passingly familiar, to political scientists, the first chapter introduces some of the key concepts, both philosophically and statistically.

1.2 Causal Inference versus Data Driven Hypothesis Generation

Existing quantitative analyses work from the theory to the data. Broadly, a dependent variable to be explained is selected. Theory, either existing or new, is used to develop an explanation, and this explanation is operationalized as an independent variable. Possible confounders are gathered. The outcome is written as a linear combination of the known confounders and the explanatory variable of interest. After accounting for the confounders, if the explanatory variable has an effect of sufficient magnitude and precision, producing a

p-value of below 0.1 or 0.05, the null hypothesis of no relationship is rejected. A relationship between the independent variable and the outcome cannot be rejected, so the causal pathway is considered significant.

The proposed methods reverse this relationship. Rather than begin with theory, I start with the data to uncover unexpected relationships. Instead of testing the effect of some pre-specified set of effects, I instead search for a small subset of predictors from a vast set of variables. The criterion for variable inclusion changes dramatically, as does the interpretation of the effects.

1.2.1 When P-values Fail

Assuming a null hypothesis of zero effect, p-values answer the question, "Within the assumed model, what is the probability of seeing an estimate at least this extreme, if there is in truth no systematic relationship between these variables?" If this p-value is sufficiently small, in that it is not likely to see such an extreme value if the true relationship were simply random noise, the estimate is deemed significant. The Bayesian critique of p-values is well established (Cohen, 1994; Efron, 1986), but I offer a different critique here.

The nature of the critique is twofold. First, p-values cannot handle the case when there are more hypotheses (k) than observations (n)$---$ even though the covariates still contain useful information. This insight is important in biostatistics, in which data often consist of microarrays, and the number of covariates can be several times larger than the sample size. Even though $k > n$, there is still information in the data that needs to be considered. This lies at the crux of several critiques of inferential methods (Brady and Collier, 2004): complex events such as revolutions and critical elections offer many more hypotheses than observations. A model with all three-way interactions among ten variables will produce 176 variables, enough to swamp even a reasonably sized data set. In the situation when k is larger than n, inferential methods fail, and these situations are not hard to find.

Second, and more relevant, as the number of k grows relative to n, each individual hypothesis is tested with less information. This is perhaps most clear in opposite terms. As n grows relative to k, each hypothesis is estimated more precisely. Even when $k < n$, once

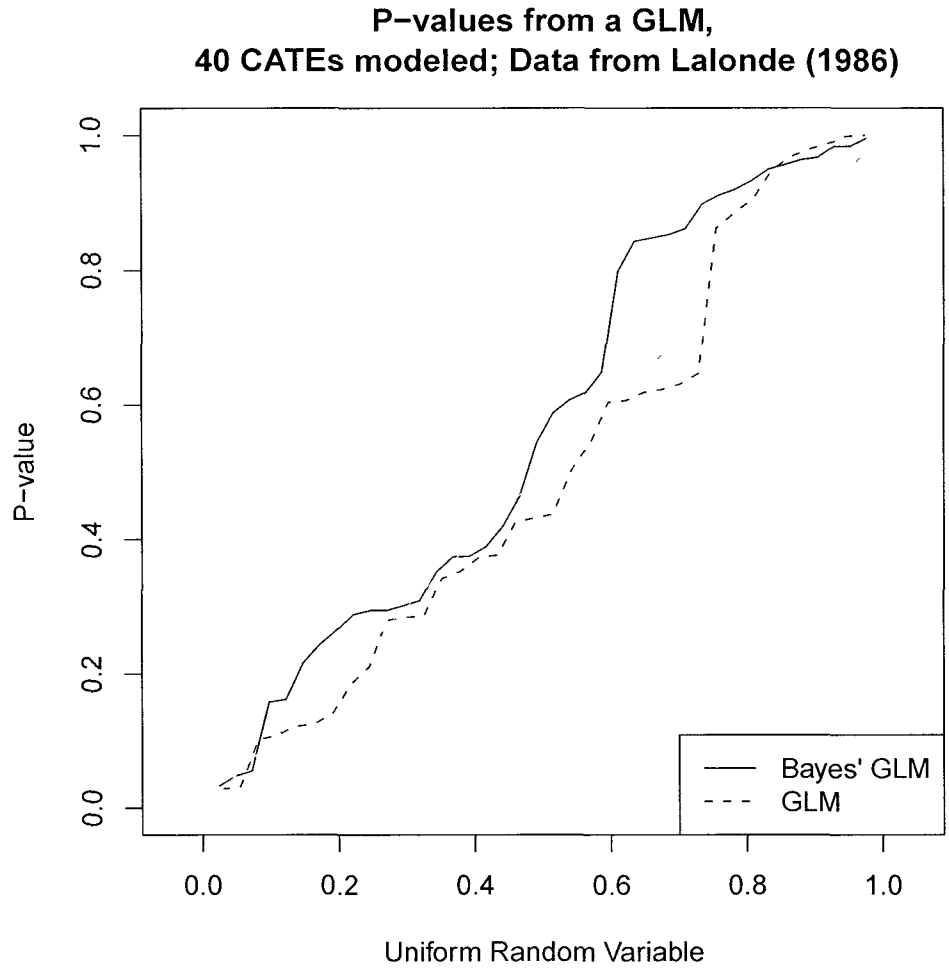

Figure 1.1: A quantile plot of p-values for CATEs from a logistic regression and Bayesian logistic regression, for the NSW data. The p-values are plotted against a uniform distribution. If the p-values were informative, they would lie below the 45 degree line. The p-values fall along this line, and are, as a group, indistinguishable from noise.

k reaches a certain point relative to n, simultaneous inference at a fixed false positive rate grows uninformative (Benjamini and Hochberg, 1995).

Consider figure 1.1, which contains a plot of p-values from estimating the probability of the National Supported Work Study recipients having a higher income after participation in the program. If the true relationship were pure noise, among a set of variables of interest, the p-values would be uniformly distributed. As a group, about 5% would be expected to be less than .05, 10% less than 0.1, and so on. Some of the covariates would be statistically

significant, but not any more than what is to be expected under random noise The model used to generate the p-values fit 79 variables to 722 observations, more than enough to be comfortably identifiable Yet, the p-values for the 40 treatment effects are indistinguishable from a uniform random variable Inference could be done some of the p-values are below 0 05, but the distribution of p-values does not generate much confidence in those results A Bonferroni correction could be used, taking 05/40 as the critical value, at which point *nothing* would be significant (for recent developments, see Benjamini and Hochberg, 1995)

In scenarios where p-values offer no guidance in variable selection, current political methodology is cast adrift It is precisely these scenarios that the proposed methods in this dissertation address

1.2.2 Data-Driven Hypothesis Generation and Model Selection

Rather than inference within an assumed model, the proposed methods identify a sparse models from within a broad class of models The fundamental tradeoff between model fit and model complexity is central to all of the proposed methods A too-complex model will overfit, performing poorly on data coming from the same generating process, a too-simple model will underfit, missing systematic relationships that would aid in prediction The approach balances fit on the observed data against model complexity, i e the number of selected variables The most complex sparse model that can be supported by the data is returned The balance is captured in an objective function that is additive in empirical loss (residual sum of squares, for example), and model complexity Reducing the empirical loss requires an increase in model complexity, and the models are fit to balance this tradeoff

Rather than a p-value criterion, I use a predictive criterion The role of effective prediction as central to positive social science has been long-established (Friedman, 1953, Maki, 2009) The predictive criterion used throughout the dissertation balances model size against model fit Adding a new variable will never worsen model fit, but, adding a new variable will increase model size The methods proposed here answer the following question "Would adding this additional variable improve prediction on a different draw of this data?" Consider, in the extreme, the case where there are as many linearly independent variables as

observations– – –the model fits the observed data perfectly; yet, on the *next* dataset, it will perform quite poorly. Balancing this tradeoff between model fit and number of variables (model dimensionality) lies at the heart of the proposed methods.

This predictive approach handles two different shortcomings of the inferential methods. First, consider the case when $k \approx n$ or $k > n$. The extent to which the most highly predictive variable reduces predictive error does not depend on how many other variables are under consideration. If the variable increases prediction error at a rate faster than at which it increases model dimensionality, it is included; else, it is not. Second, this increase in predictive power is only loosely related to its p-value. A variable with a small p-value may be substantively meaningless, but may be measured so precisely that it is deemed significant. A variable that is not significant, due to correlation with many other variables, might be an excellent predictor in the absence of some of its confounders.

Rather than testing covariates independently, the predictive criterion used through this dissertation provide a means to identify a model in which most of the parameters are assumed to be zero, balancing model fit against model size. The discovered variables may be thought of in two separate manners. From a Bayesian perspective, the selected covariates are Maximum a Posteriori estimates, assuming a Laplacian prior. From a likelihoodist perspective, the selected covariates are Best Linear Unbiased Predictors. The uncovered variables are not formally tested, in an inferential framework, but they are selected if they have sufficient explanatory power relative to their increase in model dimensionality. A more formal description of these methods follow in section 1.4.

1.2.3 A Model Selection and Variable Selection Framework

Model selection and variable selection are central to data-driven hypothesis generation, and a vast literature exists on this approach (Efron *et al.*, 2004a; Breiman, 1996; Hastie *et al.*, 2001b; Shao, 1997). Let y_i be an outcome variable of interest, $i \in \{1, 2, \ldots, n\}$. Each observation has a vector of k observed covariates, x_i, and the problem is to characterize $E(y_i|x_i)$. Let $\tilde{y}_i$ and $\tilde{x}_i$ denote unobserved draws from the same process that generated the observed data. To guard against overfitting, the methods model $E(\tilde{y}_i|\tilde{x}_i)$, as a function of

$\tilde{x}_i$. The aim is to fit the *next* dataset as well as possible, as opposed to the observed data in hand. To do this, a class of models is assumed, $\mathcal{A}$, indexed by α, and denoted $\nu_\alpha(x_i)$. Assuming squared loss, a prediction criterion is used to minimize over the class of models

$$\underset{\alpha \in \mathcal{A}}{\operatorname{argmin}} E\left((\tilde{y}_i - \nu_\alpha(\tilde{x}_i))^2\right) \tag{1.1}$$

Powerful black-box predictors have been developed that make only very weak assumptions about ν_α (Chipman *et al.*, 2010; Breiman, 2001). This performs poorly, though, if the goal is hypothesis generation. The fitted models handle arbitrary interactions, but give the researcher little guidance as to what interactions are in place. Instead, the proposed methods consider a subset of the model selection problem, that of variable selection. In this case, we assume $E(y_i|x_i) = x_i'\beta$. If we assume that most of the elements of β are 0, the model selection problem becomes one of variable selection. $\mathcal{A}$ then consists of all possible subsets of predictors in x_i, a set of size 2^k.

This is the approach taken throughout this dissertation. Each element of x_i corresponds with a hypothesis of interest to the researcher, and this set may be arbitrarily large. Variable selection then is used to identify hypotheses that we would expect to have high external validity. The application of model selection, variable selection, and data driven hypothesis generation is discussed within the context of two common empirical frameworks, that of likelihood based inference and the Neyman-Rubin-Holland causal model.

1.2.4 Model Misspecification as a Variable Selection Problem

A common, well-established manner of conducting inference is the likelihood approach. I summarize it briefly here to show where the proposed methods diverge from normal practice. Maximum likelihood estimation is most common in scenarios where the outcome variable is some limited dependent variable, such as a binary or a count variable. In these cases, the outcome is not linear in a set of covariates because the fitted values may fall outside the range of the dependent variable; fitted values may produce negative counts, or probabilities outside $[0, 1]$. Instead, a scale is found such that some transformation of the linear model stays within

the appropriate range. Intuitively, maximum likelihood estimators are the estimates that are most likely to have generated the observed data. They come with a host of positive attributes, described below.

More formally, consider an outcome variable y_i, $i \in \{1, 2, \ldots, n\}$, which is assumed to be a realization from some distribution, $f(y_i|\theta)$. I also assume $f(y_i|\theta)$ is in the exponential family, which encompasses all distributions in common usage: the Normal, the Bernoulli, the Poisson, the Negative Binomial, among others. This gives a joint distribution of the outcome as

$$P(y_1, y_2, \ldots, y_n) = f(y_1, y_2, \ldots, y_n|\theta) \tag{1.2}$$

Rather than condition on θ, the observed data is conditioned on, and likelihood function, $L(\cdot)$, is generated

$$L(\theta|y_1, y_2, \ldots, y_n) = f(y_1, y_2, \ldots, y_n|\theta) \tag{1.3}$$

It is often easier to work with the log of the likelihood, $l(\cdot) = \log L(\cdot)$. The most likely value of θ to generate the observed data y is the *maximum likelihood estimator*, $\hat{\theta}_{MLE}$.

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}}\ l(\theta|y_1, y_2, \ldots, y_n) \tag{1.4}$$

The MLE possesses several desirable properties. First, it achieves the minimum asymptotic variance (Cramer-Rao lower bound); second, it is asymptotically normal; and third, it possesses an invariance property, such that $\widehat{f(\theta)}_{MLE} = f(\hat{\theta}_{MLE})$,

Given a k-dimensional vector of characteristics for each individual, x_i, with corresponding parameters β, it is commonly assumed that the outcome of interest can be written as

$$E(y_i) = \eta(x_i'\beta) \tag{1.5}$$

$\eta(\cdot)$ is a link function, transforming the values $x'\beta$ to the scale of y. For binary outcomes, $\eta(x_i'\beta) = \frac{1}{1+exp(-x_i'\beta)}$; for count data, this becomes $\eta(x_i'\beta) = exp(x_i'\beta)$; and for normally distributed data, $\eta(x_i'\beta) = x_i'\beta$.

Regarding the broader argument, I focus primarily on the assumption in 1.2 and 1.5: that the true model is linear in some set of observed covariates. Conditioning on θ (or β, in practice), makes two subtle assumptions. First, the researcher assumes that the outcome is linear in observed covariates and that no others are necessary to characterize $f(\cdot)$. Issues over unmodeled higher-order interactions and nonlinearities are assumed away. The proposed methods address these concerns by including smoothers and a vast number of higher-order terms, as appropriate.

Second, some subset of θ is the hypothesis of primary interest. The method assumes that the only relevant hypothesis is being considered by implicitly constraining all other hypotheses to have effect zero− − −they are simply not included in the modeling process. The effect of interest, though, may vary with other covariates in systematic, important, and interesting ways, and these hypotheses are left unexplored. The proposed methods put these hypotheses back in the modeling process, by considering a broad class of hypotheses, uncovering those with the most explanatory power, and setting the remainder to zero.

1.2.5 Causal Heterogeneity as a Variable Selection Problem

Likelihood methods produce a systematic means to identify significant relationships between the outcome and an independent variable. The dominant framework for moving from a correlative relationship to a causal claim is that of the Neyman-Rubin-Holland (NRH), or potential outcomes, framework(Holland, 1986; Rubin, 1973, 1974, 1978). The NRH framework has three main components. First, it is counterfactual in nature. Each observation has a series of "potential outcomes," the outcome that would occur under any given treatmcnt. Only one of these is observed, which is the "fundamental problem of causal inference." The researcher would much rather observe each unit receiving all treatments, but only one of those worlds is actualized. Second, the treatment variable is manipulable and treatment assignment is random, so there is some positive probability put on each observation receiving

each treatment. Finally, the assignment to treatment level is assumed independent of the potential outcomes, conditional on observed covariates, so that the causal effects are not polluted by selection bias or endogeneity.

I focus here on recasting the question of causal heterogeneity as a variable selection problem. While the average treatment effect (ATE) or average treatment effect on the treated (ATT) are the most common estimands, unmodeled heterogeneity may be present. Conditional average treatment effects (CATEs), where the treatment effect varies systematically across subgroups of the data, are often left unmodeled. By characterizing a broad class of CATEs, but placing them under a variable selection constraint, subtle interactions between the treatment and recipient characteristics can be identified.

Consider a simple random sample of n observations from a population $\mathcal{P}$. Within the potential outcomes framework of causal inference (Holland, 1986), for each unit i, $Y_i(t)$ to denotes the potential value of the binary outcome that would be realized under the treatment status $T_i = t$. This notation relies upon the stable treatment unit value assumption; no interference between units and no multiple version of the treatment. Estimating causal effects requires assuming strong ignorability of treatment assignment, where the treatment level is assigned independent of potential outcomes (Rubin, 1990).

In addition, assume a treatment variable T_i is multi-valued and takes one of the $(K+1)$ possible values from the set $\mathcal{T} \equiv \{0, 1, \ldots, K\}$ where $T_i = 0$ means that unit i is assigned to the control (reference) condition. Thus, the observed outcome variable Y_i is equal to $Y_i(T_i)$. Finally, x_i to denotes an M dimensional vector of observed pre-treatment covariates for unit i where the support of this random variable is denoted by $\mathcal{X}$.

Given this setup, for each unit, the causal effect of the treatment condition $T_i = t$ (relative to the control condition $T_i = 0$) as $Y_i(t) - Y_i(0)$. The average treatment effect (ATE) is then given by,

$$\tau(t) \equiv \Pr(Y_i(t) = 1) - \Pr(Y_i(0) = 1) \tag{1.6}$$

One commonly encountered problem related to treatment effect heterogeneity is to select the most effective treatment among a large number of alternatives using the causal effect

estimates from a finite sample. That is, identifying the treatment condition t such that $\tau(t)$ is the largest, i.e., $t = \text{argmax}_{t' \in \mathcal{T}} \tau(t')$. Researchers may also be interested in identifying a subset of the treatments whose ATEs are positive. In both cases, conducting variable selection is desirable in order to avoid subsetting the data, which may lead to inefficient inference and multiple testing problems.

Another common challenge addressed is identifying subgroups of units for which a treatment is most effective (or most harmful). In other words, one wishes to identify a subset of pre-treatment covariates that efficiently characterize units to whom the treatment is most beneficial. This problem can be understood as the problem of inferring the following conditional average treatment effect (CATE) for a particular treatment condition $t \in \mathcal{T}$,

$$\tau(t; \tilde{x}) \equiv \Pr(Y_i(t) = 1 \mid \widetilde{X}_i = \tilde{x}) - \Pr(Y_i(0) = 1 \mid \widetilde{X}_i = \tilde{x}), \tag{1.7}$$

for $\tilde{x} \in \widetilde{\mathcal{X}}$ where $\widetilde{X}_i$ is a subset of the observed pre-treatment covariates X_i, and $\widetilde{\mathcal{X}}$ is its support. Since X_i is typically of a large dimension, variable selection is desirable for identifying a smaller subset of the pre-treatment covariates that are predictive of ATE.

1.3 Common Concerns with the Proposed Method, Addressed

1.3.1 Isn't this just data mining?

Yes, though I do object to the word "just." Many political scientists remain uncomfortable with using data to uncover, rather than simply test, relationships in data. It is epistemically jarring at first, but in many ways, the proposed methods can be more honest and more informative than commonly used methods. Concerns have been raised that these methods are "atheoretic." This is wrong on two counts. First, below, I describe the elegant statistical theory underlying these problems. Estimates, with known asymptotic properties are produced (Knight and Fu, 2000; Wahba, 1990; Bickel *et al.*, 2006) . They are not "data-dredging," where differing models are fit until a p-value of 0.05 or less is generated. If, by "atheoretic," it is meant that no a priori theory linking individual behavior and observable outcomes is used to justify each variable, then that is true, though this scenario is better characterized as "pretheoretic." The proposed methods search through a vast number of

variables, when we do not have the resources or ability to test each on independent datasets. Rather than asking whether a given hypothesis is significant, a large set of hypotheses can be considered. There are more hypotheses in the data than are dreamt of by current theory. These methods do not come without some losses, but the gains in terms of considering a large number of variables and fitting models of a complexity well beyond that in common practice are undeniable.

At their core, the methods presented here systematically evaluate complex models, returning parsimonious results that are easily interpretable. I am not arguing that these methods supplant traditional inference: quite the contrary. In situations where a clear hypothesis derives directly from rigorous theory, a proper model can be characterized, and, ideally, a field- or quasi-experimental dataset can be gathered, then the inferential framework is certainly appropriate. The further from this ideal, though, the more appealing the proposed methods become. I propose these methods where the researcher has a broad range of variables, is agnostic over which may be the most important, and can explain either positive or negative significant estimates of any given parameter.[1] Researchers often report "unexpected" results, those that are significant in an unexpected direction, or significant but were not anticipated by existing theory. The proposed methods provide a means to search exhaustively for these unexpected effects. These scenarios call for discovery, rather than inference. As our data grows in scope, size, and complexity, and the discipline moves from data-poor to data-rich, these methods will only grow more applicable.

1.3.2 But economists don't do it!

At conferences, two separate people have asked me if this method is yet prevalent in economics. The short answer is, not yet; citations in the economics literature consist of a single paper (Ferrari *et al.*, 2009). Economics, as a field, has developed a rigorous, comprehensive statistical structure, sometimes independent of and sometimes in collaboration with the statistics mainstream (for a fascinating exchange between the two fields, see Angrist *et al.*, 1996). It has suited the field well.

[1] Andrew Gelman has referred to these hypotheses as "vampiric" more than "empiric."

Instead of asking whether economists use these methods, I argue that political scientists absolutely should. We have many questions of arbitrary complexity, where findings of either deep complexity or deceptive simplicity could move the field forward. What correlates with the onset of war? With vote choice? With levels of social welfare expenditure?

What doesn't?

The proposed methods are within the mainstream of the "machine learning" community, which is home to a cross-section of statisticians, computer scientists, biostatisticians, industrial engineers, electrical engineers, demographers, mathematicians, business scholars, and sociologists. The community, and its corpus, is better suited to prediction rather than causal inference, though one of the proposed methods begins the work of linking the two more formally. Economists are notably poorly represented in this community, and it is well beyond the scope of this dissertation to hypothesize as to why (but see Ziliak and McCloskey, 2007). The research produced in this subfield, fitting high-dimensional models to finite data, is a vibrant, active area (for an overview, see Fan and Lv, 2009), and political science can only benefit from participating.

1.3.3 Is this inference?

No, this is not inference. Discovered results are not significant, in the normal sense. They are powerful predictors.

This has led me to refer to variable selection in this field as hypothesis generation. Rather than test simple, or obvious, hypotheses with data, the methods suggest complex hypotheses. An additional dataset, or different mode of inquiry, is necessary to establish each variable's relevance.

1.3.4 What is lost through using these methods?

The greatest loss is that of stepping outside the inferential framework. A hypothesis is generated, rather than tested; these analyses are more likely to start, rather than finish, a line of research. This should not be discounted. If a proper hypothesis, with expectations over the direction of its effect, can be produced from rigorous first principles, then the inferential

framework should be used. The methods introduced in this dissertation will only confuse the picture.

Second, these methods are relatively new and therefore under theoretical development. There is very little left to say about simple likelihood methods, such as logistic regression. The methods introduced here are an active theoretical field. It was only recently shown that the LASSO estimator is asymptotically biased (Knight and Fu, 2000). This bias generates some unexplained, yet systematic, variance, leading to the selection of improper variables that correlate with the true variables with some positive probability. In practice, LASSO estimates commonly select a non-negligible number of tiny effects. Methods that avoid this, through possessing the Oracle Property (where the estimator selects the true model with probability one as sample size grows), are multivariate versions of Hodges' estimator, and improve on the Cramer-Rao lower bound at the cost of achieving maximal risk near the variable acceptance threshold (Leeb and Potscher, 2008). Machine learning methods are not as developed as more common likelihood methods, and many questions are areas of active research.

1.4 Regularization Methods

The previous sections provided a light overview of the proposed methods, and their statistical and theoretical underpinnings. This section begins the more formal explication. Throughout this section, there is assumed an observed dependent variable for each observation, y_i $f(y_i|\beta)$, a vector of k covariates, x_i and corresponding parameters β. The log-likelihood of $f(\cdot)$ is denoted $l(\cdot)$, and the empirical loss is denoted as $J(\beta)$.

Each of these questions is addressed through recasting the estimation problem as one of "regularization." Political scientists are familiar with likelihood based methods, which minimize some form of empirical loss: least squares, logistic loss for binary outcomes, or log loss for count data. Regularization methods are a generalization of these methods, where a "penalty" is added onto the likelihood, in order to produce output with desirable properties. As an example familiar to political scientists, consider the AIC statistic, of the form

$$AIC(\beta) = -2 \cdot l(\beta) + 2 \cdot \dim(\beta) \tag{1.8}$$

A researcher intent on maximizing the log-likelihood could simply add as many linearly independent covariates as observations, producing a likelihood of one. This, of course, overfits the data and generalizes poorly. To guard against this, the empirical loss, captured by the log likelihood, is constrained, such that increases in the likelihood due to expanding the parameter space are offset by the size of the model ($\dim(\beta)$). The 2 in this statistic balances the tradeoff between model size and model fit. Different choices exist; a researcher could simply replace 2 with $\log(n)$, and arrive at the BIC statistic. The basic insight carries through to all of the methods introduced in this dissertation: model fit should be balanced against model size. Formulating this tradeoff is the hallmark of regularization methods.

Rather than constraints of the form $\dim(\beta)$, the proposed dissertation will focus on two different constraints. The first, the Least Absolute Selection and Shrinkage Operator (LASSO), produces point estimates of zero for most covariates (Tibshirani 1996). If political scientists want to consider hundreds, or even thousands, of covariates, the LASSO constraint provides a mean for selecting among them simultaneously. The LASSO has gained great traction across disciplines, from biology, where genes far outnumber the number of observations, to industrial engineering, as a means of signal processing (Hesterberg *et al.*, 2008). Political science will benefit from considering these innovations: a vote equation, or predictor of war or economic growth, can be fit that includes a near-arbitrary number of covariates.[2]

The second constraint allows for straightforward extensions to nonparametric smoothers. Smoothing splines have long been cast as a regularization method, where a set of smooth basis functions (covariates) are introduced, but a constraint is placed to balance the "curviness" of the resultant fit. This dissertation presents variable selection and smoothing under a single, regularization framework. This allows a means to select among some variables, and smooth among others, as the researcher's question dictates.

[2]The asymptotic results vary with whether or not the number of covariates grows in sample size (Shao, 1997).

Regularization methods require solving a constrained optimization, additive in an empirical loss and a positive semi-definite "penalty," of the form

$$\hat{\beta} = argmin_{\beta} \underbrace{J(\beta)}_{\text{Empirical loss}} + \underbrace{\lambda}_{\text{Smoothing Parameter}} \cdot \underbrace{\Omega(\beta)}_{\text{Penalty}} \tag{1.9}$$

Regularization methods correspond with maximum a posteriori estimates, and, in less Bayesian language, these estimators are known to be minimizers of $E(l(\beta))$(Vapnik, 2000; Scholkopf and Smola, 2001). To explain these methods, and how I use them in the proposed dissertation, I break the explanation into three components, for the loss function, penalty, and tuning parameter.

1.4.0.1 Loss Functions

The loss functions used in regularization methods are the most familiar to political scientists. The loss function is, often, a negative log-likelihood, characterizing the distance between the model and the data. For least squares regression, squared loss is used, but different losses may be used depending on the nature of the data generating process. The most commonly used loss functions are

$$\text{Squared Loss: } \sum_{i=1}^{n}(y_i - x_i'\beta)^2;\ y \in \Re \tag{1.10}$$

$$\text{Absolute Deviation: } \sum_{i=1}^{n}|y_i - x_i'\beta|;\ y \in \Re \tag{1.11}$$

$$\text{Log Loss: } \sum_{i=1}^{n} y_i x_i'\beta - \exp\left(x_i'\beta\right);\ y \in \mathbb{N} \tag{1.12}$$

$$\text{Logistic Loss: } \sum_{i=1}^{n} \log\left\{1 + \exp(y_i x_i'\beta)\right\};\ y \in \{\pm 1\} \tag{1.13}$$

$$\text{Hinge Loss: } \sum_{i=1}^{n} \max(1 - y_i x_i'\beta, 0);\ y \in \{\pm 1\} \tag{1.14}$$

$$\tag{1.15}$$

With the exception of the hinge loss, an optimal classifier studied in chapter 2, the first four losses should be familiar as negative log likelihoods in a standard maximum likelihood framework. Rather than simply minimize the empirical loss, as ML methods do, the estimators are constrained via the penalty term.

1.4.0.2 Variable Selection with the LASSO

The most commonly used variable selection method is a cross between expertise and common sense: only "relevant" variables are included, main effects are favored, and interactions or higher-order terms are only included if a strong case can be made for their inclusion. Standard data-driven variable selection methods include sequential selection methods (forwards, backwards, stagewise) and best subset methods. Sequential methods perform poorly, since a poor initial step can lead to undesirable selection afterwards. "Best subset methods" consist of evaluating all possible subsets, subject to a criterion such as Cp, AIC, or BIC. These methods underperform, since each covariate is either included in the model or not, when, a preferable model (in terms of lower prediction error, higher posterior probability, or higher penalized likelihood) would include a shrunken estimate.

Variable selection has recently been recast within a regularization framework as a penalized likelihood. In a seminal paper, Robert Tibshirani proposed the Least Angle Selection and Shrinkage Operator, or *LASSO* (Tibshirani, 1996). With y and x_i standardized, the LASSO estimator is defined as the solution to the minimization problem:

$$\hat{\beta}^{LASSO} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \sum_{j=1}^{k} |\beta_j| \tag{1.16}$$

The constraint sets some of the variables to zero, with $\lambda = 0$ giving the least squares solution and $\lambda_1 \to \infty$ returns $\hat{\beta}_{LASSO} = 0$. The algebraic intuition is most apparent within the context of an orthogonal design (i.e. $X'X$ is proportional to the identity matrix). Let $\hat{\beta}^o$ be the least-squares estimates of β and $(x)^+$ denote $x \cdot I(x > 0)$. In an orthogonal design,

the LASSO estimator can be written (see Tibshirani 1996: 269):

$$\hat{\beta}_j^{LASSO} = \hat{\beta}_j^o \left(1 - \frac{\lambda_1}{|\hat{\beta}_j^o|}\right)^+ \tag{1.17}$$

The LASSO estimator shrinks least squares estimates greater than λ_1 towards zero by factor $1 - \lambda_1/|\hat{\beta}_j^o|$. Covariates with least squares estimates less than λ_1 are estimated as zero. For non-orthogonal design, the LASSO solution proves intractable, since the penalty $\sum_{j=1}^{k} |\beta_j|$ is not differentiable at $\beta_j = 0$, although the general insights provided by the orthogonal case carries through.

In a likelihood framework, the method can be motivated out of sheer usefulness; in fact, recent algorithmic advances allow for rapid fitting with $k > n$ at the computational expense of only k least squares estimates (Efron *et al.*, 2004b). The method may also be motivated as the posterior mode, with a Laplacian prior placed over β; see Park and Casella (2008) for a fully Bayesian treatment.

The LASSO carries an informative geometric interpretation. LASSO regularization can be viewed as placing a constraint on a likelihood, with a solution where the hyperellipse $log-likelihood = k$ is tangent to the constraint. The standard form of the LASSO estimator, and a corresponding smoothed estimator,[3] is given below:

$$\hat{\beta}^{LASSO} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 \text{ subject to } \sum_{j=1}^{k} |\beta_j| \leq q_{LASSO} \tag{1.18}$$

$$\hat{\beta}^{smooth} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 \text{ subject to } \sum_{j=1}^{k} \beta_j^2 \leq q_{smooth} \tag{1.19}$$

The geometric interpretation is made clear in figure 1.2. Consider the case with only two coefficients, β_1 and β_2. In this case, the ridge constraint is the circle $\beta_1^2 + \beta_2^2 = k_2$. The LASSO constraint, in contrast, is the square $|\beta_1| + |\beta_2| = k_1$. The confidence (Scheffe) ellipse is centered at the unconstrained estimate $(\hat{\beta}_1, \hat{\beta}_2)$, and its shape is governed by $cov(\hat{\beta}_1, \hat{\beta}_2)$.

[3]This is the constraint used in random effects models, smoothing splines, ridge regression, or through assuming a normal prior over the coefficients. The resulting estimates differ in interpretation, based off whether β is assumed random or fixed, but the optimization is the same.

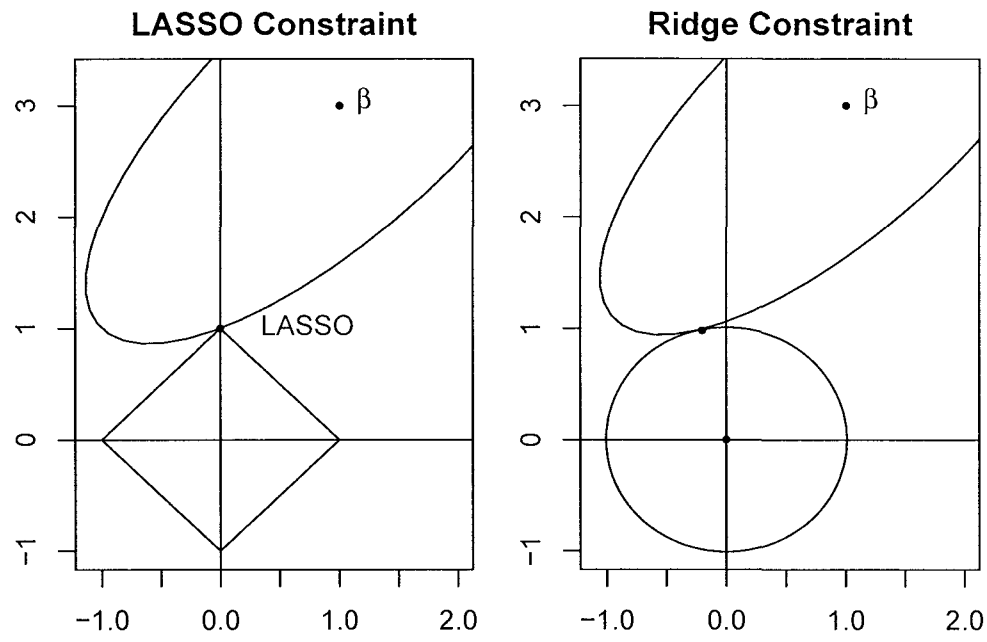

Figure 1.2: A comparison of the ridge and LASSO constraints. The LASSO constraint produces point estimates of zero, by generating point estimates at where the diamond is tangent to the ellipse.

For a given value of k_1 or k_2, the minimizer to the loss function occurs where the confidence (Scheffe) ellipse is tangent to the constraint. The ellipse will hit the smoothing constraint at a point where neither coefficient is zero. The ellipse, though, is likely to hit the square at a corner, setting some of the estimates to zero. In practice, the LASSO estimator is a powerful variable selection mechanism.

1.4.0.3 Smoothing Splines and Regularization

Often, the researcher does not know a precise functional form for the target function. It may be known to be some function of a observed variable, but little may be known about whether the effect is linear, quadratic, cubic, and so on, in the observed data. To handle this uncertainty, this dissertation accounts for the smoothness through the use of the popular nonparametric method of smoothing splines (Wahba, 1990; Gu, 2002; Shawe-Taylor and Cristianini, 2004).

In the simplest spline model, pairs of observations (y_i, t_i) are observed, with t_i having support $\mathcal{T}$. It is assumed that the systematic component of y_i is additive in a linear and a smooth component, $\eta(t_i)$. The smooth component is assumed twice differentiable function,

with $\int_{\mathcal{T}}(\eta''(t))^2 dt < \infty$. This produces a model for y_i of the form

$$y_i = d_0 + d_1 t_i + \eta(t_i) + \epsilon_i \tag{1.20}$$

The celebrated Representor Theorem of Kimeldorf and Wahba (1971) shows that the population minimizer of the form $E((y_i - \hat{y}_i)^2|t)$ can be written as $\hat{y} = R\hat{\mathbf{c}} + S\hat{\mathbf{d}}$, for $n \times 1$ vector c and 2×1 vector d. R is an $n \times n$ matrix purely determined by t and assumptions about the nature of η, while S is a low-dimensional matrix, generally linear in t. Columns in R are a series of smooth basis functions, a type of Fourier transform. Columns in S consist of an intercept and linear term for t; R is constructed so that it is uncorrelated with S. R is the penalized component, parameterizing the smooth curve, while S is the unpenalized component. With known R and S, the problem reduces to a problem of the following form:

$$\{\hat{c}_{SS}, \hat{d}_{SS}\} = argmin_{c,d}\ (\mathbf{y} - R\mathbf{c} - S\mathbf{d})'(\mathbf{y} - R\mathbf{c} - S\mathbf{d}) + \lambda_2 \mathbf{c}' R\mathbf{c} \tag{1.21}$$

Since R is an $n \times n$ matrix, the problem has more parameters $(n+2)$ than observations (n), necessitating the regularization. The level of regularization, or, in this case, smoothing, is controlled by the parameter λ_2. For $\lambda_2 = 0$, the fitted values are a complete interpolation of the data. For $\lambda_2 \to \infty$, the fitted values approach the least squares line from regressing $\mathbf{y}$ onto S, which spans the unpenalized space. Selecting λ_2 controls the balance between these two extremes. The coefficients in $\mathbf{c}$ are penalized, while those in $\mathbf{d}$ are not.

1.4.0.4 Tuning Parameter Selection

Several of the methods used here require the selection of multiple tuning parameters. To accomplish this,a GCV statistic is calculated at each fixed value of $\{\lambda_1, \lambda_2\}$ (Wahba, 1990), in order to balance model fit against model dimensionality. Given a sample size of n and model dimensionality of k, the GCV statistic is

$$GCV_{\lambda_1,\lambda_2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\beta)^2}{(1 - \frac{k}{n})^2} \tag{1.22}$$

GCV statistics are known to be inconsistent for model selection, when the model space is finite(Shao, 1997). To adjust the GCV to variable selection, I propose a Bayesian GCV (BGCV) statistic of the form

$$BGCV_{\lambda_1,\lambda_2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\beta)^2}{(1 - \frac{log(n)}{2} \cdot \frac{k}{n})^2} \tag{1.23}$$

I use the GCV when the primary problem is that of prediction. In this case, "false discoveries" are less troublesome, so long as they add some predictive power. When the goal is optimal variable selection, I use the BGCV. Simulations show similar results for GCV for BGCV, similar to differences between AIC and BIC statistics for model identification, but, as expected from theory, GCV performs slightly better on prediction, and BIC produces slightly fewer false discoveries. While I acknowledge the ad hoc nature of the BGCV statistic (for a similar use of this statistic, see Shi *et al.*, 2006), I find that it maintains a reasonable discovery rate and a low false discovery rate in both simulation and practice.

1.4.0.5 Search Strategy

The search strategy consists of a series of alternating line searches. First, λ_1 is fixed at a large value, (exp(25)). Next, λ_2 is evaluated along the set $\log(\lambda_V) \in \{-15, -14, \ldots, 10\}$, with the value producing the smallest GCV statistic selected. Given the current estimate of λ_V, λ_Z is evaluated along the set $\log(\lambda_V) \in \{-15, -14, \ldots, 10\}$, and the λ_V that produces the smallest GCV statistic is selected. We alternate in a line search between the two parameters to convergence at a given precision. After convergence at a given precision, the radius is decreased, and the precision increased. The process is repeated to a precision of .0001.

1.5 Conclusion: The Proposed Methods

Within a regularization framework, each of the proposed methods can be described rather concisely. The second chapter recasts causal heterogeneity as a variable selection problem, through the use of two LASSO constraints. One constraint is placed over pre-treatment covariates and another over the causal heterogeneity parameters of interest. This allows for small effects to be selected, even in the presence of known large effects. The third

chapter concerns the changepoint problem, where the mean function is characterized as a smooth curve with some small number of discrete breaks. A method for identifying the breakpoints, and a BIC statistic as a stopping rule is introduced. The fourth chapter is the two-dimensional version of the previous chapter. A smooth curve is fit to two dimensional data, and a series of jurisdiction-specific effects are selected with a LASSO constraint. The "mixed-penalty" method combines a LASSO constraint and a smoothing spline constraint.

The regularization framework discussed in this introduction allows for the fitting of complex models, and the LASSO constraints are used to tame the results to a reasonable number. By using the data to generate, rather than test hypotheses, a subset of predictive effects from a much larger set of possible variables can be selected, even in the presence of nonlinearities. Identifying these effects allows insight into political processes and outcomes that would not be amenable to study otherwise.

Chapter 2

Identifying Treatment Effect Heterogeneity through Optimal Classification and Variable Selection

2.1 Introduction

While much research in the causal inference literature has focused upon the overall average treatment effect, the identification of treatment effect heterogeneity plays an essential role in a number of situations that are commonly encountered by applied researchers.[1] For example, ascertaining subpopulations for which a treatment is most beneficial (or harmful) is an important goal of many clinical trials. However, the most commonly used method, subgroup analysis, is often inappropriate and remains as one of the most debated practices in the medical research community (see e.g., Assmann *et al.*, 2000; Rothwell, 2005; Lagakos, 2006).

Identification of treatment effect heterogeneity is also important for numerous other purposes. They include (1) selecting the most effective treatment among a large number of available treatments, (2) designing optimal treatment regimes for each individual or a group of individuals (e.g., Manski, 2004; Pineau *et al.*, 2007; Moodie *et al.*, 2009; Imai and Strauss, 2011), (3) testing the existence of heterogenous treatment effects (e.g., Gail and Simon, 1985; Davison, 1992; Crump *et al.*, 2008), and (4) generalizing causal effect estimates obtained from an experimental sample to a target population (e.g., Frangakis, 2009; Cole and Stuart, 2010; Hartman *et al.*, 2010; Green and Kern, 2010b; Stuart *et al.*, 2011). In all of

[1]This chapter is joint work with Kosuke Imai, Department of Politics, Princeton University.

these cases, researchers must infer how treatment effects vary across individual units with different characteristics and/or how causal effects differ across various treatments.

In this paper, we propose a method that combines optimal classification with variable selection to identify heterogeneous treatment effects when the outcome is binary. Recently, some scholars have pointed out that identification of treatment effect heterogeneity can be considered as a variable selection problem (Gunter *et al.*, 2011; Imai and Strauss, 2011). Building on this insight, Section 4.3.1 introduces the Support Vector Machine (SVM) with a separate LASSO constraint for the causal heterogeneity parameters of interest. This differs from the standard setup where a single regularization constraint is applied to all model parameters. The use of two separate LASSO constraints ensures that variable selection is performed separately for variables representing alternative treatments and/or treatment-covariate interactions. Not only are these variables different qualitatively from other variables in the model (e.g., pre-treatment covariates), but they often have relatively weaker predictive power. The proposed model avoids the ad-hoc variable selection of existing procedures by automating everything in one step (e.g., Gunter *et al.*, 2011; Imai and Strauss, 2011). The model also directly incorporates sampling weights, which are particularly useful when generalizing the causal effects estimates obtained from an experimental sample to a target population. This single step procedure contrasts with the multi-step procedures proposed in the literature (e.g., Hartman *et al.*, 2010; Green and Kern, 2010b; Stuart *et al.*, 2011).

To efficiently fit the proposed model with multiple regularization constraints, we develop an alternating line search algorithm that avoids the use of grid search, cross validation, or matrix inversion. This makes the proposed methodology more computationally efficient relative to the commonly used methods for identification of treatment effect heterogeneity such as Boosting (Freund and Schapire, 1999; LeBlanc and Kooperberg, 2010) and Bayesian Additive Regression Trees (BART) (Chipman *et al.*, 2010; Green and Kern, 2010a).

Another key advantage of the proposed methodology over Boosting and BART is that it produces a parsimonious model with a fewer number of parameters and therefore the model output can be more easily interpreted. Our model is most similar to the Bayesian logistic

regression with a non-informative prior (Gelman *et al.*, 2008). But, we use an SVM loss with two LASSO constraints rather than the logistic loss with a single Cauchy prior. This means that many of our model parameters are estimated to be zero rather than shrunk towards zero, thereby producing a parsimonious model.

To evaluate the performance of the proposed methodology, we conduct a set of simulation studies to compare it with some of the commonly used alternative methods including Boosting, BART, and Bayesian logistic regression with a non-informative prior. Our simulation studies examine the performance of the proposed method in terms of identifying how treatment effects differ across multiple treatments and how the causal effect of a treatment varies across individuals with different characteristics. As shown in Section 2.3, the results indicate that the proposed method has lower false discovery rate than the competing methods. In addition, we find that the proposed method mostly has a comparable discovery rate and competitive predictive properties to these commonly used alternatives.

For empirical illustration, we apply the proposed method to two well-known randomized field experiments from the social sciences. The results of our analysis are presented in Section 2.4. First, we analyze a get-out-the-vote field experiment where voters were randomly assigned to approximately 280 combinations, with three different appeals (civic duty, neighborhood solidarity, close election) through one of three different mobilization strategies (phone call, personal visits, and post cards) in order to ascertain their causal effects on voter turnout (Gerber and Green, 2000). We apply the proposed methodology to identify a certain combination of appeal mobilization strategies that can most effectively increase turnout. Such an analysis may help campaign managers choose the most effective mobilization strategy among a large number of possible strategies.

Second, we apply the proposed methodology to the experimental data from the National Supported Work Demonstration (NSW), which is a temporary employment program designed to help disadvantaged workers (LaLonde, 1986). The qualified workers who were assigned to the treatment group received all the benefits of the NSW program (e.g., job training and counseling). In this application, we use the proposed method to identify the groups of workers

who benefit most from the NSW program in terms of raising workers' wages. Furthermore, we show how our method can be used to generalize the causal effect estimates obtained from an experimental sample to a target population. Such an analysis helps answer the question of what impacts policy makers should expect if they were to implement the NSW program in a broader population.

Finally, Section 2.5 offers concluding remarks to summarize the contributions of the proposed methodology.

2.2 The Proposed Methodology

In this section, we describe the proposed methodology by presenting the model and developing an estimation algorithm to fit the model. We begin by formalizing the problem of identifying treatment effect heterogeneity.

2.2.1 The Framework

Consider a simple random sample of n observations from a population $\mathcal{P}$. Note that this population may not correspond directly to the target population of inference, which we denote by $\mathcal{P}^*$. Within the potential outcomes framework of causal inference (Holland, 1986), for each unit i, we use $Y_i(t) \in \{-1, 1\}$ to denote the potential value of the binary outcome that would be realized under the treatment status $T_i = t$. This notation relies upon the stable treatment unit value assumption; no interference between units and no multiple version of the treatment (Rubin, 1990). In addition, we assume that the treatment variable T_i is multi-valued and takes one of the $(K + 1)$ possible values from the set $\mathcal{T} \equiv \{0, 1, \ldots, K\}$ where $T_i = 0$ means that unit i is assigned to the control (reference) condition. Thus, the observed outcome variable Y_i is equal to $Y_i(T_i)$. Finally, we use X_i to denote an M dimensional vector of observed pre-treatment covariates for unit i where the support of this random variable is denoted by $\mathcal{X}$.

Given this setup, for each unit, we can define the causal effect of the treatment condition $T_i = t$ (relative to the control condition $T_i = 0$) as $Y_i(t) - Y_i(0)$. The average treatment

effect (ATE) is then given by,

$$\tau(t) \equiv \Pr(Y_i(t) = 1) - \Pr(Y_i(0) = 1). \tag{2.1}$$

One commonly encountered problem related to treatment effect heterogeneity is to select the most effective treatment among a large number of alternatives using the causal effect estimates from a finite sample. That is, we wish to identify the treatment condition t such that $\tau(t)$ is the largest, i.e., $t = \text{argmax}_{t' \in \mathcal{T}} \tau(t')$. Researchers may also be interested in identifying a subset of the treatments whose ATEs are positive. In both cases, conducting variable selection is desirable in order to avoid subsetting the data, which may lead to inefficient inference and multiple testing problems.

Another common challenge we address in this paper is to identify subgroups of units for which a treatment is most effective (or most harmful). In other words, one wishes to identify a subset of pre-treatment covariates that efficiently characterize units to whom the treatment is most beneficial. This problem can be understood as the problem of inferring the following conditional average treatment effect (CATE) for a particular treatment condition $t \in \mathcal{T}$,

$$\tau(t; \tilde{x}) \equiv \Pr(Y_i(t) = 1 \mid \widetilde{X}_i = \tilde{x}) - \Pr(Y_i(0) = 1 \mid \widetilde{X}_i = \tilde{x}), \tag{2.2}$$

for $\tilde{x} \in \widetilde{\mathcal{X}}$ where $\widetilde{X}_i$ is a subset of the observed pre-treatment covariates X_i, and $\widetilde{\mathcal{X}}$ is its support. Since X_i is typically of a large dimension, variable selection is desirable for identifying a smaller subset of the pre-treatment covariates that are predictive of ATE.

We next turn to the description of the proposed model that combines optimal classification and variable selection to identify treatment effect heterogeneity. For the remainder of the paper, we assume the strong ignorability of treatment assignment (Rosenbaum and Rubin, 1983),

$$\{Y_i(0), Y_i(1), \ldots, Y_i(K)\} \perp\!\!\!\perp T_i \mid X_i = x \quad \text{and} \quad 0 < \Pr(T_i = t \mid X_i = x) < 1 \tag{2.3}$$

for all $t \in \mathcal{T}$ and $x \in \mathcal{X}$. This assumption is guaranteed to hold in randomized experiments and is also common when estimating causal effects in observational studies. Under this

assumption, equations (2.1) and (2.2) reduce to the following,

$$\tau(t) = \Pr(Y_i = 1 \mid T_i = t) - \Pr(Y_i = 1 \mid T_i = 0) \tag{2.4}$$

$$\tau(t; \tilde{x}) = \Pr(Y_i = 1 \mid T_i = t, \widetilde{X} = \tilde{x}) - \Pr(Y_i = 1 \mid T_i = 0, \widetilde{X} = \tilde{x}), \tag{2.5}$$

respectively. Thus, as shown below, we use optimal classification and variable selection within the context of regression modeling.

2.2.2 The Model

In modeling treatment effect heterogeneity, we begin by considering the following linear classification rule for the binary outcome variable Y_i,

$$\widehat{c}_i = \operatorname{sgn}(\widehat{Y}_i) \tag{2.6}$$

$$\widehat{Y}_i = (\hat{\mu} + Z_i^\top \hat{\beta} + V_i^\top \hat{\gamma}) \tag{2.7}$$

where Z_i is an L_Z dimensional vector of covariates that represent treatment effect heterogeneity and V_i is an L_V dimensional vector containing the rest of the covariates in the model. For example, if researchers wish to identify the most efficacious treatment condition among all the possible treatments, then Z_i would consist of $(K+1)$ indicator variables, each of which represents a different treatment or control condition whereas V_i would include pre-treatment variables that need to be adjusted for within the model. Similarly, if identifying subgroups of units for which a treatment is most beneficial (or harmful), Z_i would include variables representing interactions between the treatment indicator variable and all the pre-treatment covariates of interest. In this case, V_i would include all the main effects with respect to the pre-treatment covariates. The idea here is to separate the causal heterogeneity variables of interest from the rest of the variables.

In estimating the coefficients, β and γ, we adapt the support vector machine classifier (SVM) and place separate LASSO constraints over each set of coefficients (Vapnik, 1995; Tibshirani, 1996; Bradley and Mangasarian, 1998; Zhang, 2006). Our model differs from the standard model by allowing β and γ to have separate LASSO constraints. This is motivated

by the qualitative difference between the two sets of parameters, and also by the fact that often causal heterogeneity variables have weaker predictive power than other variables.

Specifically, define the "hinge-loss" function as $|x|_+ = max(x, 0)$. We formulate the support vector machine as a penalized squared hinge-loss objective function (Wahba, 2002), with two separate l_1 constraints to generate sparsity in the covariates,

$$\sum_{i=1}^{n} w_i \cdot |1 - Y_i \cdot (\mu + Z_i^\top \beta + V_i^\top \gamma)|_+^2 + \lambda_Z \sum_{j=1}^{L_Z} |\beta_j| + \lambda_V \sum_{j=1}^{L_V} |\gamma_j|, \tag{2.8}$$

where λ_Z and λ_V are separate LASSO penalty parameters for β and γ, respectively, and w_i is an optional sampling weight, which may be used when generalizing the results obtained from one sample to a target population.

Our objective function is similar to several existing LASSO variants but there exist important differences. For example, the elastic net introduced by Zou and Hastie (2005) places the same set of covariates under both a LASSO and ridge constraint to help reduce mis-selections among correlated covariates. In addition, the group LASSO introduced by Yuan and Lin (2006) groups different levels of the same factor together so that *all* of a factor is selected, rather than particular levels and rotational invariance is preserved. In contrast, the proposed method places separate LASSO constraints on the qualitatively distinct groups of variables so that variable selection is performed among causal heterogeneity parameters of interest.

2.2.3 The Estimation Algorithm

The estimation algorithm progresses in three steps: the data are rescaled, the model is fit for a given value of λ_Z and λ_V, and each fit is evaluated using a generalized cross validation statistic.

2.2.3.1 Rescaling the Covariates

LASSO regularization requires rescaling covariates (Tibshirani 1996). Following standard practice, all pre-treatment main effects are centered and given standard deviation one. Higher order terms are interactions between the lower-order standardized terms.

We model two different forms of causal heterogeneity. In the first, treatments consist of multiple crossed factors, where each individual receives exactly one level of several different factors. For example, there may be three different treatments, with four treatment conditions each. The main treatment effects consist of the twelve columns of indicator variables (three times four). Each of these columns is left uncentered, keeping most of their entries to zero, but given standard deviation one. Interaction treatment effects are constructed as the product of these lower order terms.

The second form of causal heterogeneity we consider is that of a single treatment interacted with multiple pre-treatment covariates, in order to ascertain for which subgroups a treatment is most efficacious. In this case, standardization occurs in three steps. First, the pre-treatment covariates is standardized, but left uncentered. The uncentered, standardized treatment indicator is then interacted with the pre-treatment covariates. Each treatment/pre-treatment covariate interaction covariates is then centered on the treated units, and the untreated observations are set to zero.

2.2.4 Fitting the Support Vector Machine

The support vector machine is estimated through a series of iterated LASSO fits, using two simple observations. First, for a given outcome $Y_i \in \{\pm 1\}$, $|1 - Y_i\hat{Y}_i|_+^2 = (Y_i - \hat{Y}_i)^2 \cdot \mathbf{1}(1 \geq Y_i\hat{Y}_i)$, which allows the SVM to be written as a least squares problem on a subset of the data. Second, for a given value of $\{\lambda_Z, \lambda_V\}$, rescaling Z and V allows the objective function to be written as a LASSO problem, with a tuning parameter of 1, as

$$\sum_{i=1}^{n} w_i \cdot |1 - Y_i \cdot (\mu + Z_i^\top \beta + V_i^\top \gamma)|_+^2 + \lambda_Z \sum_{j=1}^{L_Z} |\beta_j| + \lambda_V \sum_{j=1}^{L_V} |\gamma_j| = \tag{2.9}$$

$$\sum_{i=1}^{n} w_i \cdot |Y_i - (\mu + \frac{1}{\lambda_Z} Z_i^\top (\lambda_Z \beta) + \frac{1}{\lambda_V} V_i^\top (\lambda_V \gamma))|_+^2 + \sum_{j=1}^{L_Z} |\lambda_Z \beta_j| + \sum_{j=1}^{L_V} |\lambda_V \gamma_j| \tag{2.10}$$

These two observations allow the problem to be coerced into a form that can be fit by an efficient LASSO algorithm (Efron *et al.*, 2004b). The algorithm alternates between

estimating a model on the subset of observations $\{i|1 \geq Y_i\hat{Y}_i\}$, and then re-estimating this set of active observations. We describe the algorithm in greater detail immediately below.

To begin the algorithm, a value of $\{\lambda_Z, \lambda_V)$ is selected. The data consist of a binary outcome $Y_i \in \{\pm 1\}$, $i \in \{1, 2, \ldots n\}$, an $n \times L_Z$ matrix of causal heterogeneity covariates Z with associated parameters β, and an $n \times L_V$ matrix of pre-treatment covariates V with associated parameters γ. Initialize the coefficients and fitted values $[\hat{\beta}^0_{\lambda_Z}|\hat{\gamma}^0_{\lambda_V}] = \vec{0}$, $\hat{\mu}^0 = 0$, and $\hat{Y}^0_i = 0\ \forall i$.

Let $\mathcal{A}^{(t)}$ denote the set of all active observations at iteration t, the set for which $\{i|1 \geq Y_i\hat{Y}^{(t-1)}_i\}$. This is the set of observations to which the LASSO model is fit. Initialize $\mathcal{A}^0 = \{i|1 \geq Y_i\hat{Y}^0_i\}$, which, by construction, is all observations at initialization. Let $X_{\mathcal{A}^{(t)}}$ denote the submatrix of the $n \times k$ matrix X, consisting of all rows of X that are in $\mathcal{A}^{(t)}$; let $\widetilde{X}$ denote the matrix X with columns centered. The algorithm progresses in six steps:

1. Generate the submatrices for the design matrix and outcome as

$$ZV^{(t)} = \left[\frac{1}{\lambda_Z}Z_{\mathcal{A}^{(t-1)}} \,\middle|\, \frac{1}{\lambda_V}V_{\mathcal{A}^{(t-1)}}\right] \tag{2.11}$$

$$Y^{(t)} = Y_{\mathcal{A}^{(t-1)}} \tag{2.12}$$

2. Estimate the LASSO coefficients as

$$[\hat{\beta}|\hat{\gamma}]_{LASSO} = \underset{[\beta|\gamma]}{\operatorname{argmin}} \sum_{i \in \mathcal{A}^{(t-1)}} (\tilde{Y}^{(t)}_i - \widetilde{ZV}^{(t)\prime}_i[\beta|\gamma])^2 + \sum_{j=1}^{L_Z}|\beta_j| + \sum_{j=1}^{L_V}|\gamma_j| \tag{2.13}$$

3. Update $[\hat{\beta}^{(t)}|\hat{\gamma}^{(t)}] = \frac{1}{2} \cdot [\hat{\beta}^{(t-1)}|\hat{\gamma}^{(t-1)}] + \frac{1}{2} \cdot [\hat{\beta}|\hat{\gamma}]_{LASSO}$

4. Update the intercept as the difference in means between Y and the fitted values, with respect to the active observations as

$$\hat{\mu}^{(t)} = \bar{Y}^{(t)} - \overline{\{ZV^{(t)}[\hat{\beta}^{(t)}|\hat{\gamma}^{(t)}]\}} \tag{2.14}$$

5. Calculate the current fitted values for all observations as

$$\hat{Y}_i^{(t)} = \hat{\mu}^{(t)} + \left[\frac{1}{\lambda_Z}Z \,\middle|\, \frac{1}{\lambda_V}V\right]_i' [\hat{\beta}^{(t)}|\hat{\gamma}^{(t)}] \tag{2.15}$$

6. Update the active set as $\mathcal{A}^{(t)} = \{i|1 \geq Y_i\hat{Y}_i^{(t)})$

For a fixed value of $\{\lambda_Z, \lambda_V\}$, steps (1)-(6) are iterated to convergence. The coefficients $[\hat{\gamma}|\hat{\beta}]$ are rescaled at the end to their original scale.

2.2.5 External criterion

The algorithm produces coefficient estimates for a given value of $\{\lambda_Z, \lambda_V\}$. This section describes both the statistic we use to assess fit, as well as the search strategy implemented to identify the tuning parameters.

2.2.5.1 External Criterion

At each fixed value of $\{\lambda_Z, \lambda_V\}$, we calculate a GCV statistic (Wahba, 1990), in order to balance model fit against model dimensionality. Define n as the sample size, n_0 as the sample size of observations in $\mathcal{A}$ at convergence, and $\hat{Y}_i$ as $\hat{Y}_i^{(t)}$ at convergence. The number of non-zero coefficients provides an unbiased estimate of the dimensionality of a LASSO model (Zou *et al.*, 2007), so we take as our criterioin the GCV statistic

$$GCV_{Y,\hat{Y};\lambda_Z,\lambda_V} = \frac{\frac{1}{n}\sum_{i=1}^{n}|1 - Y_i\hat{Y}_i|_+^2}{(1 - \frac{k}{n_0})^2} \tag{2.16}$$

2.2.5.2 Search Strategy

Our search strategy consists of a series of alternating line searches. First, we fix λ_Z at a large value, (exp(25)), effectively setting all causal heterogeneity parameters to zero (Osborn, Presnell, and Turlach). Next, λ_V is evaluated along the set $\log(\lambda_V) \in \{-15, -14, \ldots, 10\}$, with the value producing the smallest GCV statistic selected. Given the current estimate of λ_V, λ_Z is evaluated along the set $\log(\lambda_V) \in \{-15, -14, \ldots, 10\}$, and the λ_V that produces the

smallest GCV statistic is selected. We alternate in a line search between the two parameters to convergence at a given precision. After convergence at a given precision, the radius is decreased, and the precision increased. The process is repeated to a precision of .0001.

2.3 Simulation Studies

In this section, we conduct two simulation studies to evaluate the performance of the proposed method relative to the commonly used alternatives: Boosting (adaboost as implemented in R package `ada`), BART (as implemented in R package `bayestree`), and Bayesian logistic regression with a non-informative prior (as implemented in R package `arm`). The first set of simulations corresponds to the situation where researchers are interested in selecting the most effective treatment among a large number of possible treatments. The second set of simulations considers the case where we wish to identify subpopulation of units for which a treatment is most effective. In both cases, we assume that the treatment variable T_i is independent of the observed pre-treatment covariates X_i. This assumption holds in randomized experiments or in certain observational data where covariate balance is achieved via matching or other procedures. Finally, we examine four different sample sizes, $n \in \{500, 1000, 2500, 5000\}$, in both sets of simulations. For each scenario, we run 1000 simulations.

2.3.1 Identifying the Best Treatment

We begin by presenting the simulation results for selecting the best treatment among a large number of available treatments. We use two settings, one with correct model specification and the other with misspecified models, where unmodeled nonlinear terms are added to the data generating process.

In the simulations with correct model specification, we have one control condition, forty-nine additional distinct treatment conditions, and three pre-treatment covariates. Using our notation, this means that Z_i consists of fifty treatment indicator variables and V_i represents a vector of three pre-treatment covariates plus an intercept, i.e., $L_Z = 49$ and $L_V = 4$. Among the forty-nine treatments, three of them have substantive average effects whose magnitude

is approximately equal to 7, 5, and -3 percentage points, respectively. The remaining 46 treatment indicator variables have non-zero but negligible average effects where effect sizes are within ± 1 percentage point. All pre-treatment covariates, on the other hand, are assumed to have substantive predictive power.

We independently sample the pre-treatment covariates from a multivariate normal distribution with mean zero and a randomly generated covariance matrix. Specifically, an $L_V \times L_V$ matrix, $U = [u_{ij}]$, was generated with $\sqrt{u_{ij}} \sim norm(0, 1)$ and the covariance matrix is given by $U^\top U$. The design matrix for the forty nine treatment variables is orthogonal and balanced. The true values of the coefficients are set as $\beta = \{7.5, 3.3, -2, \ldots\}$ and $\gamma = \{50, -30, 30\}$ where the "..." denotes 47 remaining coefficients drawn from a uniform distribution on $[-0.7, 0.7]$. Finally, the outcome variable $Y_i \in \{-1, 1\}$ is sampled according to the following probability,

$$\Pr(Y_i = 1 \mid Z_i, V_i) \quad = \quad a(Z_i^\top \beta + V_i^\top \gamma + b) \tag{2.17}$$

where an affine transformation defined by constants $\{a, b\}$ is applied such that the magnitude of the ATEs roughly equals the values specified above.

For the simulations with incorrectly specified models, we include unmodeled nonlinear terms based on the pre-treatment covariates in the data generating process. Specifically, V_i includes the interaction term between the first and second pre-treatment covariates and the square term of the third pre-treatment covariate in addition to the main effect term for each of the three covariates, i.e., $L_V = 5$. As before, the outcome variable is generated after an affine transformation in order to keep the size of the ATEs approximately equal to the pre-specified levels given above so that the two sets of simulation results can be compared.

Figure 2.1 summarizes the simulation results. Under each simulation scenario, we compute both the false discovery rate (FDR) and the discovery rate (DR) for each method. The first row of the figure presents the FDR with respect to the largest estimated effect. That is, we compute the proportion of times the largest estimated effect is actually not the true largest effect. The second row shows the corresponding DR, which represents how often a method can correctly identifies the largest effect as the largest. The results show that across

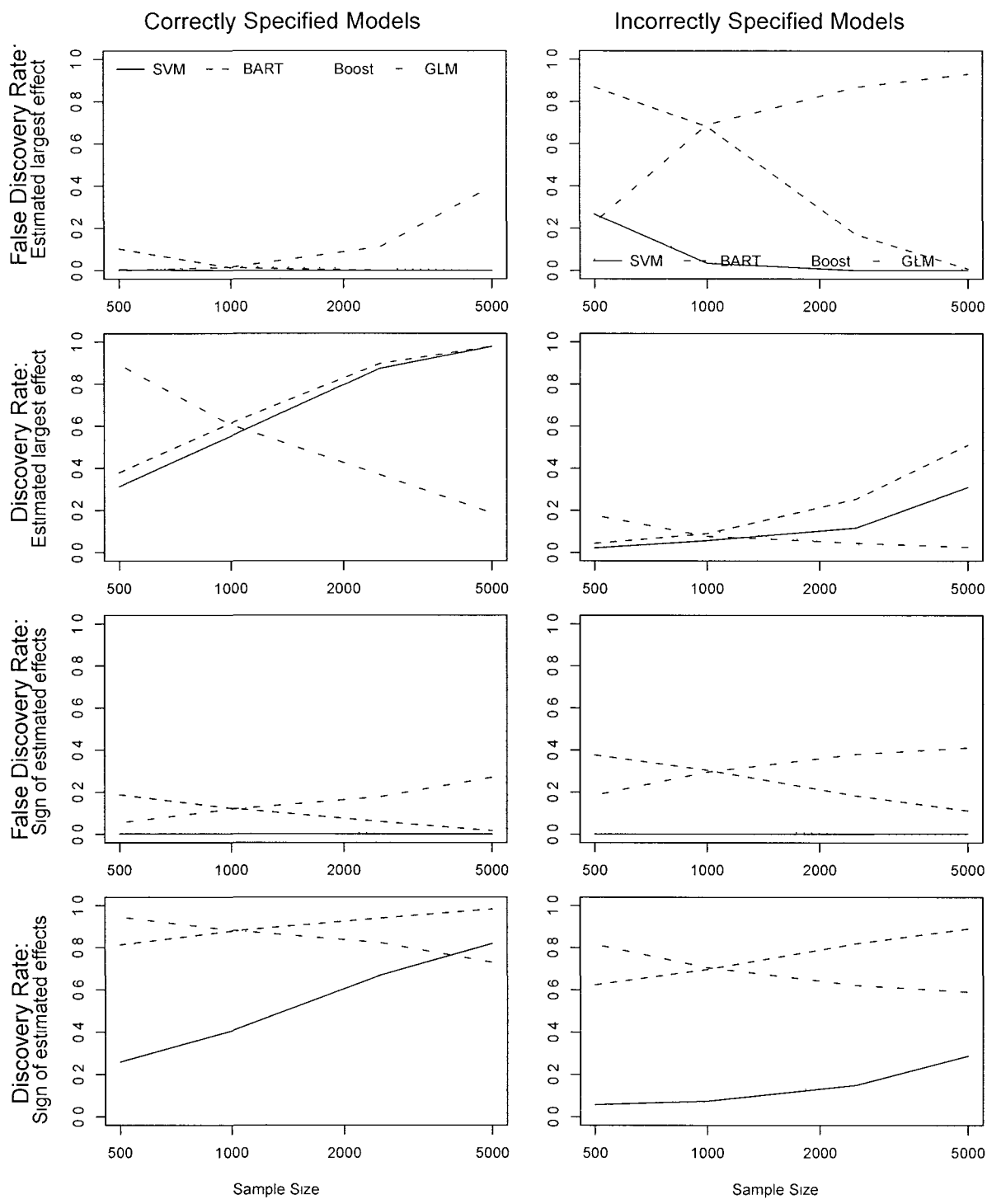

Figure 2.1: False Discovery Rates and Discovery Rates for Selecting the Best Treatments among a Large Number of Available Treatments. Simulation results with correct specification (left column) and incorrect specification (right column) are shown. The figure compares the performance of the proposed method (`SVM`; solid lines) to that of BART (`BART`; dashed lines), Boosting (`Boost`; dotted lines), and Bayesian logistic regression with a non-informative prior (`GLM`; dashed-dotted lines). The top row presents how often the largest estimated effect is actually not the true largest effect. The second row shows how often a method can correctly identifies the largest effect as the largest. The third row plots how often a method identifies the sign of estimated non-zero effects incorrectly, while the fourth row presents the proportion of sign agreement for the three treatments with

simulations the proposed method (SVM; solid lines) has small FDR while its DR is competitive with other methods. In particular, the proposed method dominates all other methods in terms of FDR when the model is correctly specified. The comparison with BART reveals a key feature of the proposed method. The latter less often identifies the best treatment, but when it does the method does so more accurately. As expected, the performance of the proposed method improves as the sample size increases though this is not necessarily the case for some other methods. For all methods, model misspecification increases FDR and reduces DR.

The final two rows of Figure 2.1 present FDR and DR for the sign agreement of treatments with substantive effects (those treatments whose ATEs are 7, 5, and -3 percentage points). The third row plots how often a method identifies the sign of estimated non-zero effects incorrectly. The fourth row presents the proportion of sign agreement for the three treatments whose ATEs are of substantive magnitude. Similar to the results above, the proposed method has small FDR across various simulation scenarios. However, the proposed method is conservative in that its DR is lower than some of the alternative methods considered here. The comparison with BART most clearly illustrates this point. As before, the performance of the proposed method improves as the sample size increases and if the model is specified correctly.

2.3.2 Identifying Subpopulations for Which a Treatment is Beneficial

In the second set of simulations, we consider the problem of identifying subpopulations for which a treatment is beneficial (or harmful). In this case, we are interested in identifying interactions between a treatment and observed pre-treatment covariates. The key difference between this simulation and the previous one is that in the current setup causal heterogeneity variables (treatment-covariate interactions) may be correlated with each other as well as other non-causal variables. In contrast, the previous simulation setting assumes that causal heterogeneity variables (treatment indicators) are independent of each other and other variables.

In the current simulation, we have a single treatment condition, i.e., $K = 1$, and twenty pre-treatment covariates X_i. The pre-treatment covariates are all based on the multivariate normal distribution with mean zero and a random variance-covariance matrix as in the previous simulation study although in this simulation five of them are discretized using 0.5 as a threshold. In our setting, causal heterogeneity variables Z_i consist of twenty treatment-covariate interactions plus the main effect for the treatment indicator while V_i is composed of the main effects for the pre-treatment covariates. As a result, we have $L_Z = 21$ and $L_V = 20$

Given this setup, we generate the outcome variable Y in the same way as in Section 2.3.1 according to the linear probability model. There are two pre-treatment covariates that interact with the treatment in a systematic manner. We apply an affine transformation so that an observation whose values for these two covariates are one standard deviation above the mean have the conditional average treatment effect of approximately 4 and -2.5 percentage points. Specifically, we set $\beta = \{2.5, -1.5, \ldots\}$ and $\gamma = \{50, -30, 30, 20, -20, \ldots\}$ where the $\ldots$ denotes uniform draws between $[-0.7, 0.7]$.

In the left column of Figure 2.2, we compare false discovery rate (FDR) and discovery rate (DR) of the largest effect for our proposed method (`SVM`; solid lines) with those for Bayesian logistic regression with a non-informative prior distribution (`GLM`; dotted and dashed lines). The right column of the figure gives the same plots for non-zero substantive effects. For Bayesian GLM, we consider two rules; one based on posterior means of coefficients (dashed lines) and the other selecting coefficients that are statistically significant with p-values below 0.1 (dotted lines). The interpretation of these plots is identical to that of the plots in Figure 2.1. Unlike the simulations given in Section 2.3.1, neither BART nor boosting provide a simple rule for variable selection in this setting and hence the results are not reported in this figure. Figure 2.2 shows that when compared with the Bayesian GLM, the proposed method has a lower FDR for both estimated largest effect and substantive effects. The GLM with the p-value thresholding yields an FDR that is closer to the FDR of the proposed method, but the latter produces a higher DR and hence is more powerful.

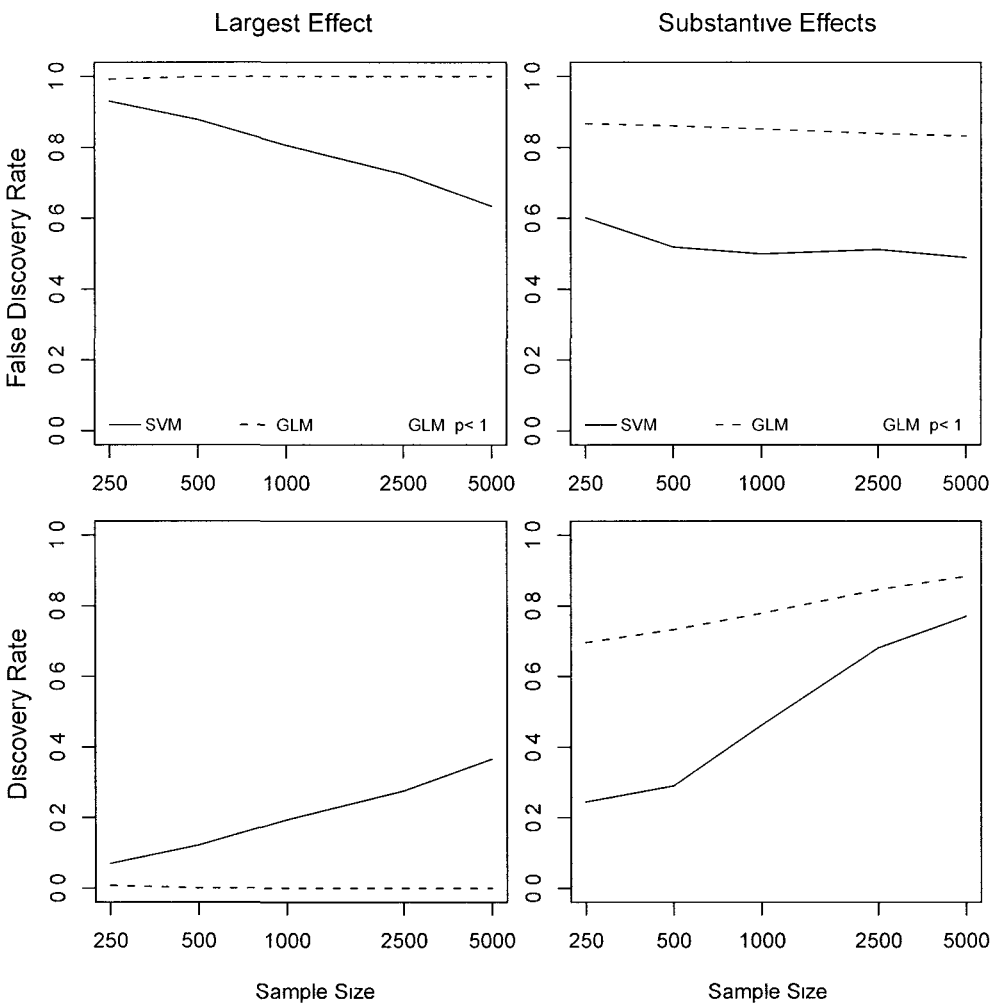

Figure 2.2: False Discovery Rates and Discovery Rates for Identifying Subpopulations for Which a Treatment is Most Effective (or Harmful). The figure compares the performance of the proposed method (SVM; solid lines) with the Bayesian logistic regression with a non-informative prior (GLM; dashed and dotted lines). For Bayesian GLM, we examine the estimates based on posterior means (dashed lines) and the statistical significance (p-value less than 0.1). The top left plot presents how often the largest largest estimated effect is actually not the true largest effect. The bottom left plot shows how frequently a method can correctly identifies the largest effect as the largest. Similarly, the right column shows the plots about FDR and DR with respect to substantive effects.

To further evaluate the relative performance of the proposed method in this simulation setting, we consider the classification rule based on each method. We then apply these classification rules to a new simple random sample of 2000 observations from the same data generating process and then compute two types of payoffs for assigning the treatment to an observation. First, the "probability payoff" p_i for assigning the treatment to observation i

is calculated as $p_i = \Pr(Y_i(1) = 1 \mid X_i) - \Pr(Y_i(0) = 1 \mid X_i)$. The probability payoff is the extent to which administering the treatment makes the event $Y_i = 1$ more likely. We next define the "classification payoff" as $c_i = 2 \times \mathbf{1}\{p_i > 0\} - 1$. That is, the classification payoff is 1 if the treatment makes observing $Y_i = 1$ more likely, and -1 if it makes observing $Y_i = 1$ less likely. For each observation treated, payoff p_i or c_i is received while for untreated observations, payoff zero is received. The classification rule for each method is to treat if $\hat{p}_i > 0$.

Finally, we compute the cumulative classification and probability payoffs by considering the situation where only a certain subset of the new sample can be classified to the treatment group. This addresses the possibility that policy makers can only afford giving the treatment to a certain number of units because of a budget constraint. The cumulative payoffs for the maximum $k\%$ possible treated units can be computed by ordering all the units according to the estimates of p_i and then classifying no greater than the $k\%$ top units with positive estimated payoffs.

Figure 2.3 evaluates the relative performance of the proposed method in terms of cumulative classification (left column) and probability (right column) payoffs. The horizontal axis represents the maximum percentage of new observations that can be classified to the treatment condition. Each row represents different sample sizes for simulations. As the benchmarks, we also include the random classification rule as well as the oracle classification rule where the oracle knows each true p_i, and treats only those observations with positive p_i.

The figure shows that, in terms of prediction, our proposed method is competitive with others. As expected, the performance of all the methods approaches that of the oracle classification rule as the sample size increases. The method dominates other methods, in terms of both cumulative probability and classification payoff. Unlike other methods, the proposed method's cumulative payoffs do not decrease sharply, with payoffs eventually plateauing as the maximum percentage increases and approaches 100%. This indicates that the classification rule based on the proposed method is conservative in terms of identifying observations

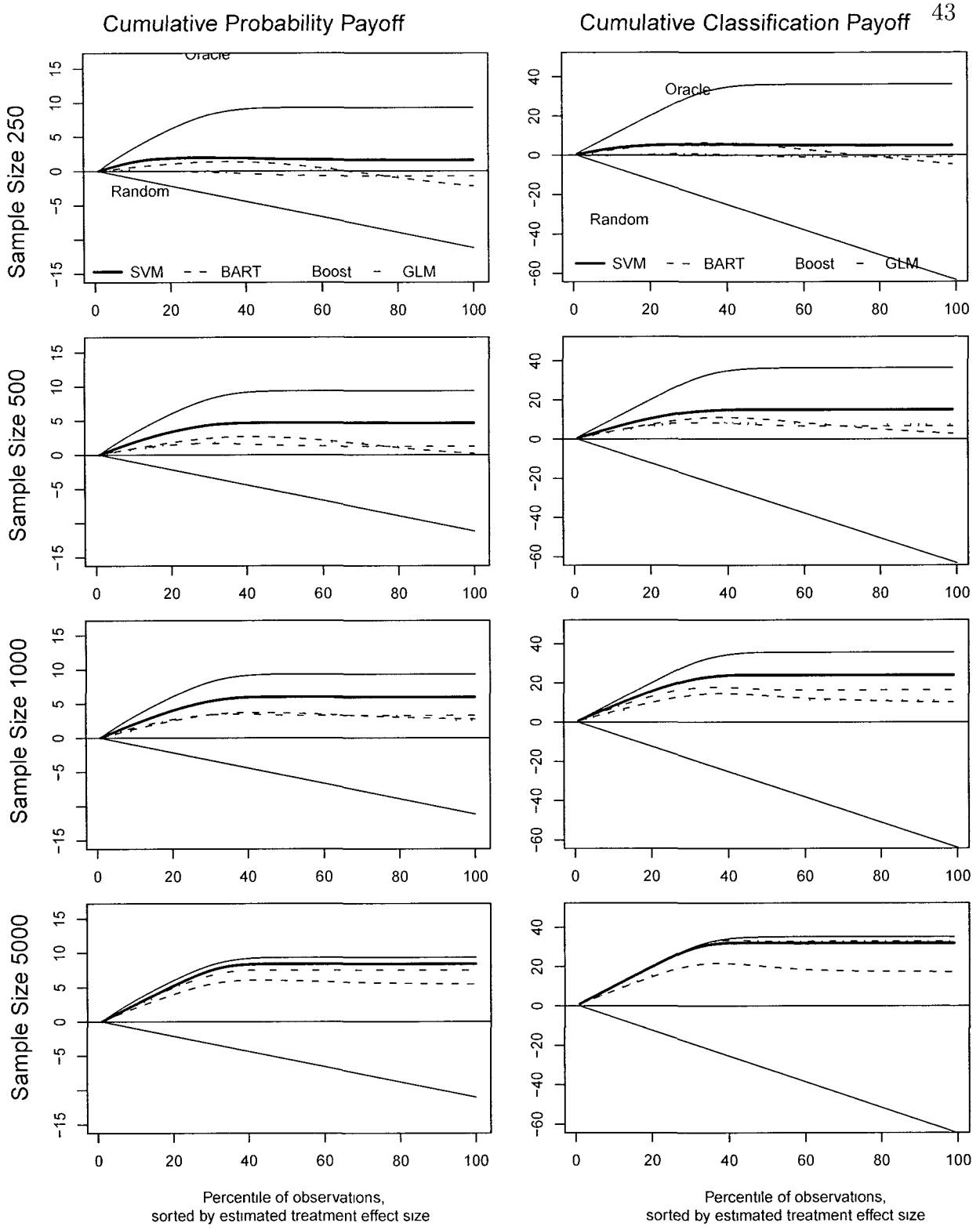

Figure 2.3: Comparison of Cumulative Classification and Probability Payoffs across Methods. The horizontal axis represents the maximum percentage of new observations that can be classified to the treatment condition. The proposed method (`SVM`; thick solid lines) is compared with BART (`BART`; dashed lines), Boosting (`Boost`; dotted lines), and the Bayesian logistic regression with a non-informative prior (`GLM`; dash-dotted lines). The plots also include the oracle classification rule and the random classification rule (thin solid lines) as the benchmarks. Each row represents different sample sizes for simulations.

that benefit from the treatment but rarely classifies observations to the treatment when the treatment is harmful for them.

This important advantage of our method is shown more clearly in Figure 2.4. In this figure, we examine the rate of change of the cumulative classification payoff (the left column of Figure 2.3), decomposed into its positive and negative components. The left column shows the proportion of units assigned to the treatment that actually benefit from the treatment, while the middle column shows the proportion of those units that are hurt by the treatment. The oracle never misclassifies observations and hence is represented by the horizontal line at zero in the figures of the middle column. The right column presents the total classification payoff at each percentile, i.e., the positive effects (left column) minus the negative effects (middle column). Each row represents a different sample size.

Figure 2.4 shows that when the sample size is small, the proposed method has the advantage of selecting more observations which benefit from the treatment than those who are harmed by it. In contrast, other methods often incorrectly classify observations to the treatment even when they are harmed by the treatment. This can be seen from the figures in the middle column where the result based on the proposed method (`SVM`; solid thick lines) stays close to the horizontal zero line when compared to other methods. Similarly, in the right column, the results based on the proposed method stay above zero. When these lines go below zero as they do for other methods, it implies that a majority of observations assigned to the treatment are worse off by receiving the treatment. The disadvantage of the proposed method is its conservativeness. This can be seen in the left column where at the beginning of the percentile the solid thick line is sometimes below other methods, for smaller sample sizes. As the sample size increases, and the advantage of the proposed method pesists.

2.4 Empirical Applications

In this section, we apply the proposed method to two landmark field experiments in the social sciences. First, we analyze the get-out-the-vote (GOTV) field experiment where forty nine unique combinations of mobilization techniques were randomly administered to registered New Haven voters in the 1998 election (Gerber and Green, 2000). Second, we

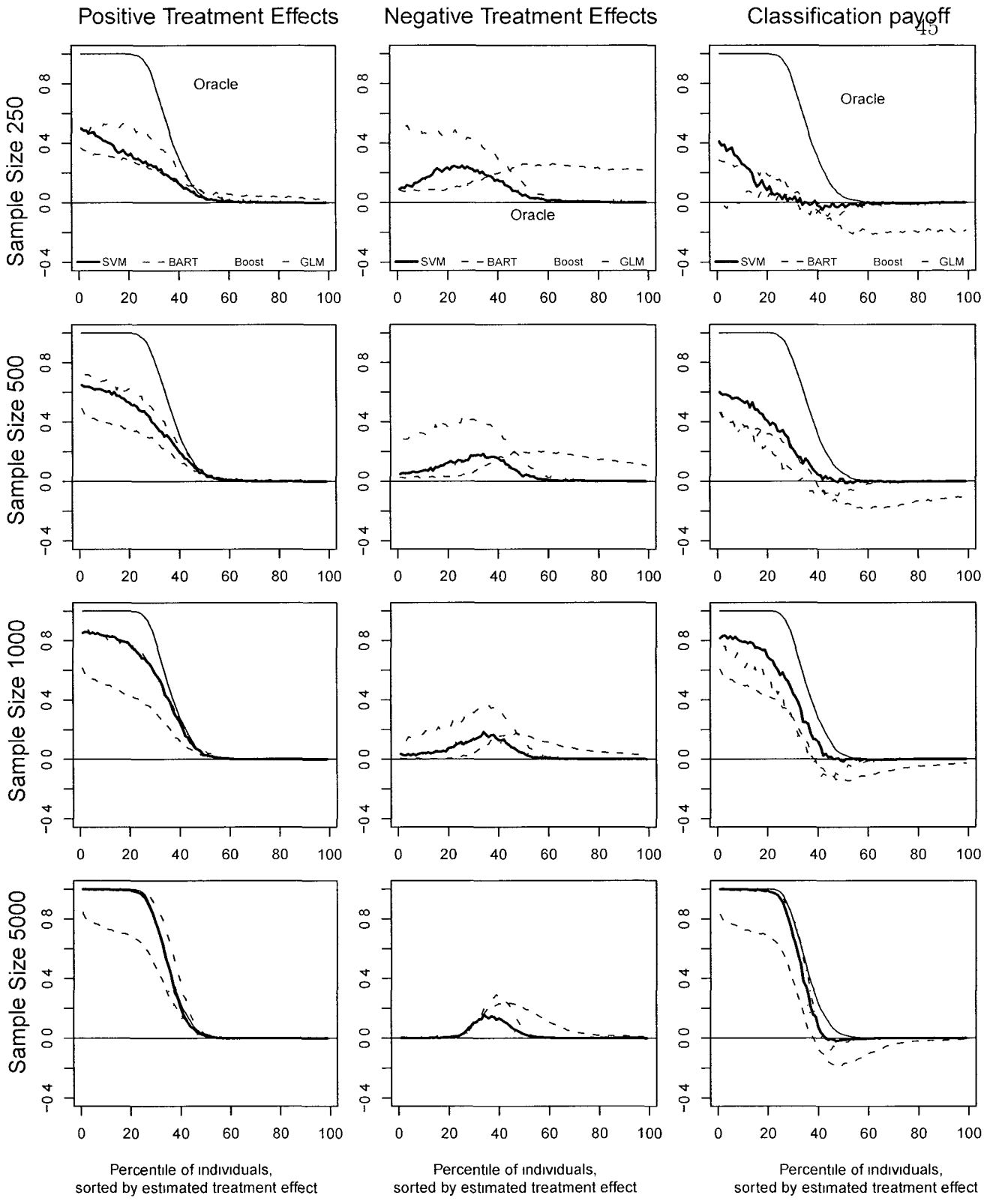

Figure 2.4: Rate of Change in the Classification Payoff for Each Method. The figure presents the proportion of treated units (based on the classification rule of each method) who benefit from the treatment (left column), are harmed by the treatment (middle column), and the difference between the two (right column) at each percentile of the total sample who can be assigned to the treatment. The oracle (solid lines) never misclassifies the observations and hence is identical to the horizontal line at zero in the middle column. The proposed method (SVM; solid thick lines) makes fewer misclassification than other

analyze the data from the National Supported Work Demonstration (NSW) where the job training program was randomly assigned to qualified disadvantaged workers. In both cases, we use the proposed method to identify treatment effect heterogeneity.

2.4.1 Selecting the Best Get-Out-the-Vote Mobilization Strategy

We first analyze the New Haven GOTV field experiment. To avoid the problem of possible interference between voters, we focus on 14,774 voters in single voter households. For the purpose of illustration, we also ignore the implementation problems documented in Imai (2005) and analyze the most recent data set. The original experiment used an incomplete, imbalanced factorial design, with four factors consisting of: one of three appeals (civic duty, neighborhood solidarity, or a close election), zero to three mailings sent, seven possible phone messages, and a personal visit. The control group consists of 5,269 voters. Additional information on each voter includes age, residence ward, whether registered for a majority party, whether she voted in the 1996 election, and whether she abstained in the 1996 election. All main-, two-, three-, and four-way interactions generate 279 combinations. The number of observations assigned to the treatment combinations range dramatically, with eleven combinations having only a single observation and a maximum of 7,424 (receiving at least one mailing).

We apply the proposed method to select the best GOTV mobilization strategy out of 279 alternatives by identifying treatment combinations that have non-zero effects. We consider two model specifications. In Model 1, the causal heterogeneity variables Z includes the binary indicator variables of 279 treatment combinations, i.e., $K_Z = 279$. In Model 2, we interact these binary treatment indicators with the past turnout, i.e., $K_Z = 558$. For both models, we have the same set of the non-causal variables V, which consist of the main effect terms of three pre-treatment covariates, there two-way interaction terms among these variables, and the square of age variable, i.e., $K_V = 11$.

Table 2.1 present the estimated non-zero coefficients for the models with (right column) and without (middle column) the interactions between the treatments and the turnout indicator variable for the 1996 election. As shown in the table, the proposed methodology

				Treatment Model	Previous voter Model
Pretreatment Covariate Coefficients					
			Intercept	-0.1280	-0.1282
			Age	0.0065	0.0066
			Majority Party	0.0617	0.0619
			1996 Voter	0.2040	0.2162
			1996 Voter * Majority Party	0.1002	0.1018
			1996 Abstained	-0.2328	-0.2323
			1996 Abstained*Age	-0.0038	-0.0038
			1996 Abstained*Majority Party	-0.0206	-0.0201
CATE Coefficients					
Treatment Schedule					
Visited	Phoned	Mailings	Appeal		
No	No	Any	Any	-0.0240	-0.0231
No	No	Two	Civic Duty	-0.0028	-0.0025
No	No	Two	Close Election	-0.0034	-0.0032
No	Yes	One	Civic Duty	-0.0147	-0.0119
No	Yes	One	Civic Duty	-0.0078	-0.0064
No	Yes	Two	Civic Duty	0.0160	0.0149
No	Yes	Two	Civic Duty	-0.0364	-0.0352
No	Yes	Three	Civic Duty	-0.0245	-0.0195
No	Yes	Three	Solidarity	-0.0201	-0.0105
No	Yesl	Two	Close Election	0.0063	0.0056
Yes	No	No	Solidarity	0.0033	0.0020
CATE Coefficients for Previous Voters					
No	Yes	One	Civic Duty	·	-0.1211
No	Yes	Three	Civic Duty	·	-0.0269
No	Yes	One	Solidarity	·	-0.0068
No	Yes	Any	Solidarity	·	0.0323
Yes	No	No	Any	·	0.0060

Table 2.1: Estimated Non-zero Coefficients for the Models With and Without Interactions Between Treatments and Turnout in the 1996 Election. The coefficients can be read based off of the treatment schedule. For example, the first CATE coefficient is an estimated effect for someone who was not visited, not phoned, and received any mailing or appeal. Estimated coefficients of the treatment variables have been rescaled so that they correspond to the estimated Conditional Average Treatment Effect.

produces a small set of non-zero coefficients and estimates all other coefficients to be zero.

Model 1 (the model without the interaction terms) shows that among 279 possible treatments combinations including a personal visit are the most efficacious. Every negative effect corresponds with a treatment that does not contain a personal visit. While a personal visit increases turnout on average by about 2-3 percentage points, phone calls and mailings alone do not appear to increase turnout.

	Sample Average	Estimated ATE, Treatment Model	Estimated ATE, Previous voter model
Personal Visit	0.0389	0.0250	0.0235
Phone Call	-0.0459	0.0044	0.0043
Mailing	-0.0025	-0.0037	-0.0035

Table 2.2: Estimated average treatment effect, for each personal visits, phone calls, and mailings. The sample average appears in the leftmost column. The next two columns contain the estimated ATEs from the two models fit using the proposed method. The sharp negative effect for the phone call disappears, while the positive effect for a personal visit is estimated at a substantively important level.

To ease interpretation, the fitted models were used to estimate ATEs for each treatment type; the results are presented in table 2.2. Personal visits have the strongest impact on turnout. The finding that a phone call depresses turnout has sparked a debate in the literature (Imai, 2005). The selected model predicts only a negligible impact for from phone calls, suggesting that the strong negative effect is the result of imbalance in the original design, rather than representative of a systematic effect.

Table 2.1 suggests underlying complex relationships within the data. There are several coefficients that vary by whether the individual had voted in the past. Among recipients of a phone call who were not visited, the number of mailings and the type of appeal appear to be interacting. Table 2.3 illustrates this complex interaction, where each cell contains the average number of voters by treatment type, among those individuals who were called but not visited. The Close Election appeal provides the starkest example. Increasing the number of mailings from zero to three with a Close Election appeal discourages previous non-voters (22% to 9%), but encourages previous voters (59% to 76%). Civic Duty appeals

grow less effective for previous non-voters, as the number of mailings increases (19% to 6%), but there is no effect on previous voters. The Solidarity appeal has little impact on previous non-voters, but discourages previous voters (62% to 48%).

The proposed method provides several insights into this experiment that have gone unnoticed previously. These insights can directly inform GOTV planners. It is well known that visiting potential voters is the most reliable way to increase turnout; it is also the most expensive. In lieu of canvassing voters, Close Election appeals should be made to previous voters, and followed up with mailings. Continued mailings discourage previous non-voters, but a phone call with a Close Election appeal encourages them. The Civic Duty and Solidarity appeals do not encourage turnout.

2.4.2 Identifying Workers for Whom a Job Training Program is Beneficial

Next, we apply the proposed methodology to the Manpower Demonstration Research Corporation's National Supported Work (NSW) Program, which was conducted from 1975 to 1978 over 15 sites in the United States. Disadvantaged workers who qualified for this job training program consisted of welfare recipients, ex-addicts, young school dropouts, and ex-offenders. Participants were unemployed and had not maintained a job for more than three months of the past half year. The job training was randomly administered to $3,214$ such workers while $3,402$ belonged to the control group. Our analysis focuses upon the subset of these individuals previously used by other researchers (LaLonde, 1986; Dehejia and Wahba, 1999). In this reduced sample, the size of the treatment and control groups is 297 and 425, respectively. We consider the binary outcome of interest measured as whether the earnings increased after the job training program (1978) compared to the earnings before the program (1975). The pre-treatment covariates include 1975 earnings, age, years of education, race (black, white, or hispanic), marriage status (married or single), whether a worker has a college degree, and whether the worker was unemployed in 1975.

In our analysis, we use the proposed methodology to answer two important questions related to treatment effect heterogeneity. First, we seek to identify the subpopulations for

which the job training program is beneficial. The program was administered to the heterogeneous group of workers and hence it is of interest to investigate whether the treatment effect varies as a function of individual characteristics. Second, we show how to generalize the results based on this experiment to a target population. Such an analysis is important for policy makers who wish to use experimental evidence when deciding whether to implement this job training program in a certain population. For illustration, we generalize the experimental results to the 1978 Panel Study of Income Dynamics (PSID), which oversamples low-income individuals. Within this PSID sample, we focus on those who had been unemployed in the previous year in order to avoid severe extrapolation. This subsample is labeled PSID-2 in Dehejia and Wahba (1999).

In our model, the matrix of non-causal variables, V, consists of thirty nine pre-treatment covariates. These include the main effects of age, years of education, and the log of one plus 1975 earnings, as well as binary indicators for race, marriage status, college degree, and whether the individual was unemployed in 1975. We also use square terms for age and years of education, and every possible two-way interactions among the pre-treatment covariates are included. The matrix of causal heterogeneity variables Z includes the binary treatment and interactions between this treatment variable and each of the thirty-nine pre-treatment covariates. This yields $K_Z = 40$ and $K_V = 39$.

Using this model specification, we conduct two separate analyses. First, we fit the model to the NSW experimental sample to identify the subpopulations of workers for which the job training program is beneficial. Second, we generalize these results to the PSID sample in order to estimate the ATE for these low-income individuals. Table 2.4 shows the marginal distribution of covariates. The differences across three samples are quite substantial. The PSID respondents are older, better educated, and more likely to be married and have a college degree than NSW participants. The proportion of blacks in the PSID sample is much greater than in the NSW experimental sample. In addition, PSID respondents earned more income, on average, than NSW participants. All differences, except for proportion hispanic, are statistically significant at the 5% level.

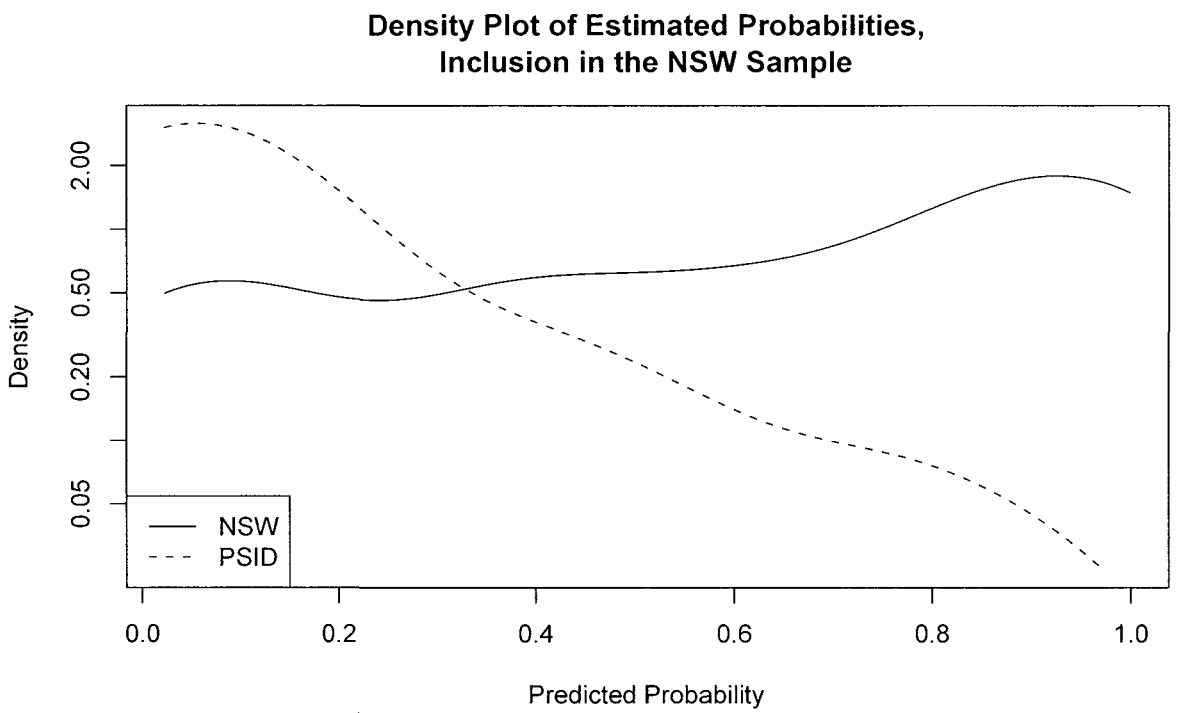

Figure 2.5: Density plot of estimated probabilities used to generate probability weights for extrapolation from the NSW sample to the PSID sample. The two samples have dramatically different distributions.

Finally, when generalizing the results from the NSW experimental sample to the PSID sample, we incorporate the sampling weights in the proposed methods. Unfortunately, sampling weights are not available in the original data and hence, for the purpose of illustration, we construct them by fitting the Bayesian logistic regression with a non-informative prior using the V matrix as the predictors. We then take the inverse estimated probability of being in the NSW sample as the weights used in the proposed method. Figure 2.5 shows the diistribution of the estimated probabilities for each sample. The two samples differ greatly, and the proposed method uncovers different conditional effects for each sample.

Table 2.5 presents the results. The first row shows the estimated ATE for each sample whereas the remainder of the table presents the estimated (non-zero) additional marginal effects above the ATE for each sample. For example, in the NSW sample, the estimated ATE for whites is $0.0240 = 0.0415 - 0.0175$, whereas that for a married worker who was unemployed for 1975 and whose age is two years below the average can be calculated as $0.0307 = 0.0415 + .0409 - 2 \times 0.0284$.

In analyzing the NSW experimental sample (left column), several basic sets of results appear. The average treatment effect is estimated at 4.15%. This is commensurate with the difference in means estimate, 6.84%, and the least squares estimate after controlling for the pre-treatment covariates in V, 6.03%; both estimates are significant at the 10% level.

In terms of heterogeneity, the program was more effective for blacks and Hispanics, though these effects were offset for blacks without a degree and hispanics with a higher previous income. Married participants fared better. For participants unemployed through 1975, the treatment was more effective for those who were older or had more education. For a policymaker interested in designing a future program, characterizing effect heterogeneity in previous studies can be very useful.

Extrapolating from the NSW to the PSID sample reveals little in the way of a treatment effect, most likely due to the differences between the two samples, as shown in figure 2.5. There is no estimated average treatment effect, but the proposed method predicts that

2.5 Concluding Remarks

Identification of treatment effect heterogeneity is essential for answering common questions in both scientific research and policymaking. The proposed method has been shown effective in selecting the best treatment from a large number of possible treatments, identifying individuals for whom a treatment is most efficacious, and in generalizing from the sample to a different population of interest.

Three central insights were made. The first was separating out the two qualitatively distinct sets of covariates, pre-treatment covariates and causal heterogeneity covariates. In common scenarios, the pre-treatment covariates are much more influential than the treatment. A variable selection technique that ignores this will unduly favor pre-treatment covariates over the treatment effects. Second, while recent work has fit "black-box" models to estimate heterogeneity, we argue for the use of models that are interpretable. Unlike the existing methods like Boosting and BART, the proposed method yields a parsimonious model by selecting a small number of parameters that characterize treatment effect heterogeneity of interest. The resulting model is easier to interpret and the associated estimates have low

false discovery rate and reasonable discovery rate. Finally, we continue recent work that has started bringing machine learning methods to questions of causal inference. Recasting the causal heterogeneity problem as one of variable selection links the insights of both subfields.

Appeal Type		Number of Mailings Sent 0	1	2	3
Close Election	Non-Voter, 1996	0.22	0.12	0.14	0.09
	Voter, 1996	0.59	0.66	0.67	0.76
Civic Duty	Non-Voter, 1996	0.19	0.17	0.10	0.06
	Voter, 1996	0.54	0.52	0.58	0.58
Solidarity	Non-Voter, 1996	0.10	0.10	0.08	0.17
	Voter, 1996	0.62	0.56	0.56	0.48

Table 2.3: Estimated probabilities of voting in 1998, for the subset of individuals who were not visited but were called by phone. The impact of the appeal varies dramatically with whether the individual voted previously.

Variables	NSW	PSID
age	24.52	36.09
years of education	10.27	10.77
black	0.8	0.39
hispanic	0.11	0.07
married	0.16	0.74
no college degree	0.78	0.49
earning in 1975	3042.76	7568.66
Sample size	722	253

Table 2.4: Sample Means of Pre-treatment Covariates for the NSW Experimental Sample, and the 1978 Panel Study of Income Dynamics (PSID) Sample.

Quantities of interest	NSW	PSID
Average treatment effect (ATE)	0.0415	0.0000
Additional main marginal effects above the ATE		
One additional year in squared age	−0.0003	0.0000
One additional year of squared education	0.0038	0.0000
Married	0.0459	0.0000
White	−0.0175	0.0000
Additional interactive marginal effects above the ATE		
Black, no degree	−0.1006	0.0000
Hispanic, one ppt increase in 1975 income	−0.0270	0.0000
One additional year of education, unemployed	0.0283	0.0000
One additional year of age, unemployed	0.0105	0.0000
Black, One ppt. increase in 1975 income	0.0047	0.0000
Married, one ppt increase in 1975 income	0.0000	0.0056

Table 2.5: Estimated Heterogeneous Treatment Effects on the Probability that the Job Training Program Increases Earnings. Estimates are given separately for the NSW experimental sample and the 1978 Panel Study of Income Dynamics (PSID) sample. The estimated average treatment effect (ATE) for each sample is given in the first row. The rest of the table presents the estimates of additional marginal effects above the ATE. For example, in the NSW sample, the estimated ATE for whites is $0.0240 = 0.0415 - 0.0175$, whereas that for a married worker who was unemployed for 1975 and whose age is two years below the average can be calculated as $0.0307 = 0.0415 + .0409 - 2 \times 0.0284$.

Chapter 3

Finding Jumps in Otherwise Smooth Curves: Identifying Critical Events in Political Processes

3.1 Introduction

Political processes are generally stable and smooth, but occasionally explosive around critical events (Pierson, 2004)[1]. Simultaneously identifying the jumps and estimating a smooth curve poses a particular statistical challenge. In this paper, we adapt a smoothing spline in a manner that allows identification of both the location and number of breaks in a time series, producing a fitted function that represents both critical and secular change.

Our interest in this problem began when we noticed the "9-11 problem" while trying to fit smoothers to George W. Bush's approval ratings, as illustrated in figure 1. Smoothers, to some extent or another, show an uptick in Bush's approval well before 9-11. The spline in figure 3.1 would lead to the conclusion that the increase in approval began in mid-July, which is clearly incorrect. Bush's approval was slowly trending downwards until 9/10, and then by 9/12 it was up around 80 percent, or even higher. Before 9/11, the spline estimate systematically overestimates Bush's approval, while systematically underestimating Bush's approval post 9/11.

Similarly, when using a loess smoother, optimal choice of the tuning parameter resulted in a curve nearly identical to the generalized smoothing spline, also missing 9/11. Decreasing the span until it picks up the 9/11 jump creates too much variance elsewhere. From a time series standpoint, while Presidential approval may be described well *on average* as an AR1

[1]This chapter is joint work with KevinEng , Department of Statistics, University of Wisconsin-Madison.

process (Erikson *et al.*, 2002), it is not *everywhere* the same process. The "memory" in the process immediately post-9/11 was rather short, while during more mundane times, the memory may be longer. Nathaniel Beck and colleagues, using a Kalman filter, insert a jump at 9/11, which does leads to a far better fit (Beck *et al.*, 2006). Inserting the jump at 9/11, though, assumes the answer to the question we are asking. While admitting 9/11 as a break seems sensical from even a casual glance at the data, an equally strong case could be made for a wide variety of events, such as the invasion of Iraq or Afghanistan, the events surrounding Hurricane Katrina, and so on. We simply do not know when to stop adding jumps.

This brought to the fore the two questions central to our method. Given data with a time component (or any natural ordering), we develop a method that identifies both the location and number of jumps, while fitting a smooth curve elsewhere. The method fits a smoothing

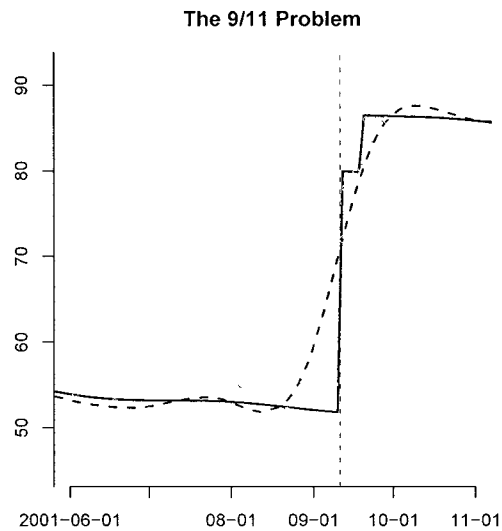

Figure 3.1: Smoothing splines are dashed and our method is solid. The smoothing splines show an uptick in Bush's popularity in mid-July, well before 9/11.

spline to the data, while using a binary segmentation algorithm to sequentially add "jumps" to the spline's unpenalized space. We develop a modified BIC statistic as a stopping rule.

We use the method here on two different sets of observed data. The first is George W. Bush's approval ratings across multiple pollsters between January 25, 2001 and October 24, 2007. Our algorithm picks out two critical events in Bush's term, plus or minus a day or two: 9/11/2001 and the invasion of Iraq. We also analyze the median of the first dimension DW-NOMINATE, both by Congress and by party. With the DW-NOMINATE medians, we detect jumps in the Congressional median between 1910-1912 and 1874-76, most likely corresponding to the rebellion against Joseph Cannon and the end of Reconstruction. We find a jump between 1930-1932 for the Democratic party, corresponding to the beginning of the New Deal. We find a jump for the non-Democratic party in 1818-1820, which is most likely a false positive.

The method provides several advances over existing methods. While there are several standard methods for finding a structural break in a parameter (as in Calderia and Zorn 1998, e.g.), ranging from the simple Chow test to Bayesian methods that estimate the location of the change-point (Western and Kleykamp, 2004), most do not offer a technique for finding both the location and *number* of breaks. Recent work by Arthur Spirling has addressed this issue explicitly, through the use of reversible jump Markov Chain Monte Carlo, in finding "turning points" in civilian casualties through the Iraq war (Spirling, 2007b; Green, 1995). In contrast to Spirling, we embed our method within a nonparametric framework through the use of smoothing splines. The method employed by Spirling allows for identification of breaks in a given parameter in a structural model, while we search for a series break in the mean function itself, rather than any particular parameter. We are able to avoid heavy parametric assumptions about either the systematic or random component of the model, which allows extension to a broad array of spline models and semiparametric models (Gu, 2002).

A rather straightforward change in loss function could easily allow for modeling limited dependent variables, similar to methods that search for breaks in parameters in generalized

linear models (Spirling, 2007a). This would allow for the identification of change-points within limited dependent variables. Finally, to ease interpretation, our method returns a serial order of breaks and a corresponding modified BIC statistic. This allows the researcher to present the jumps in terms of importance, as well as flexibility in the stopping decision.

3.2 Methods

3.2.1 A brief review of smoothing techniques.

Our interest with the presidential polling data is in summarizing its progression over time in a clear way while accounting for noisy measurements taken at irregular times. Any summarization is a tradeoff between an estimate that is too variable and too complex to be readily interpretable and one that is too smooth and glosses over too much information. We search in between for a parsimonious, intelligible, de-noised estimate.

In the interest of producing an interpretable, data driven estimate of the mean function, researchers familiar with polling data will be familiar with loess smoothing techniques for summarizing the data (Cleveland, 1979). A generalization of the weighted average, the loess balances the fit between a smooth curve and no smoothing by a bandwidth parameter.

Selection, or estimation, of this bandwidth parameter is a common characteristic of many smoothing methods. Recent work by Luke Keele has introduced a political science audience to the issues involved with semi- and non-parametric smoothing (Keele 2006, 2008), and we refer the reader there for background.

We wish to highlight the use of smoothing splines for obtaining a good functional estimate. We rely here on the functions in the `gss` library in `R` (Gu, 2002; R Development Core Team, 2008). In this paper, we use the REML algorithm to estimate the smoothing parameter throughout (Krivobokova and Kauermann, 2007).

Given data (y_i, x_i), we assume that y_i is linear in a smooth function of x_i and a mean-zero error term that is independently, identically distributed with finite fourth moment. We also

assume that the function is sufficiently "smooth," in that $\int\{f''\}^2 dt < \infty$.

$$y_i = f(x_i) + \epsilon_i \tag{3.1}$$

$$E(\epsilon_i) = 0 \tag{3.2}$$

$$var(\epsilon_i) = \sigma_\epsilon^2 \tag{3.3}$$

The cubic smoothing spline fits a function that has two components. The first, unpenalized component is the linear trend to the data. The second, penalized component allows for a nonlinear fit. Controlled by a smoothing parameter, λ, the estimator is a compromise between the least squares line ($\lambda \to \infty$) and complete interpolation of the data ($\lambda = 0$). The cubic smoothing spline minimizes the penalized residual sum of squares:[2]

$$\hat{f}_{SS} = min_f \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int \{f''\}^2 dt \tag{3.4}$$

The second component is the penalty on the "non-linear" part. If f is the sum of a linear part and a smooth (C^∞) non-linear part, for example by admitting a Taylor expansion, then the linear part of the function will have second derivative zero, and hence is not considered in the penalty. The smoothing parameter λ controls the tradeoff between smoothness and the least-squares regression line. The smoothing spline is widely applied and software packages implementing cubic smoothing splines are common (Gu, 2002; Wahba, 1990).

3.2.2 A different kind of function.

These methods are applicable when we believe the true function underlying our data is a smooth function measured with some noise. The 9/11-problem, though, highlighted an aspect of Bush's approval data that the smoothing spline could not handle. We are interested in a function that is everywhere smooth except for a small number of discontinuous jumps. We develop a method for identifying discontinuous shifts in the mean function, fitting a single smooth function but modeling breaks in the intercept. Discontinuities of this form

[2]We formulate the problem in a least squares framework, but distributional assumptions could be added so as to characterize the data-generating process (See Gu 2002 for an extensive discussion with examples). This would result in a penalized likelihood rather than penalized regression.

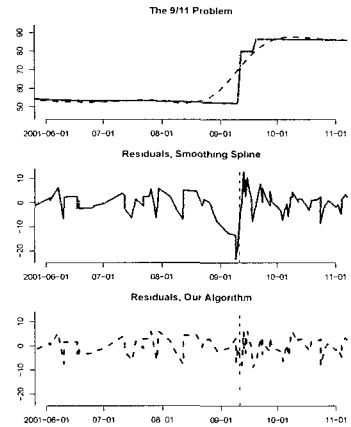

Figure 3.2: Smoothing splines are dashed and our method is solid. The smoothing splines show an uptick in Bush's popularity in mid-July, well before 9/11.

represent a problem for smoothers. They can be viewed as a form of model misspecification or omitted variable bias. The smoother, when it smooths over the jumps, will leave a distinct residual pattern: residuals will be systematically above on one side of the jump and below on the other. Figure 3.2 illustrates the residual pattern around 9/11 for both the smoothing spline and a smoothing spline with a partial spline added at 9/11.

This misspecification creates a problem for selecting the smoothing parameter. Consider the jumps as sustained shifts in the mean from one part of the function to another. Residual variance estimates will be overstated, by adding part of the "jump-to-jump" variance to the true residual variance.

If we knew the location of jumps in a function, we might add them directly to the unpenalized space of the spline. This would augment the unpenalized space, resulting in a partial spline (Gu, 2002). If $\phi(x) = I\{x > \text{"9/11"}\}$, then the partial spline is the minimizer:

$$\hat{y}_{PS} = \arg\min_{f,\beta} \sum_{i=1}^{n} (y_i - f(x_i) - \beta\phi(x_i))^2 + \lambda \int \{f''(t)\}^2 dt \tag{3.5}$$

That is, known functions are not penalized and are accounted for in the fit directly. One might think of them as being initially subtracted out, and the remainder smoothed.

"Breaks" in the residuals can be detected by eyeball and automatically. We suggest the application of binary segmentation procedures (Sen and Srivastava, 1975), which look for breaks in a constant mean function.

Let $\hat{e}_1, \ldots, \hat{e}_n$ be the estimated residuals from the smoothing spline fit, after removing the linear trend. Define the partial sums $S_i = \hat{e}_1 + \ldots + \hat{e}_i$. Under the assumption of homogeneity and known variance, Sen and Srivastava derive the uniformly most powerful test for a breakpoint. They show that the most likely breakpoint can be found at the maximal t statistic, $Z = max_{1 \leq i \leq n} |Z_i|$ for

$$Z_i = \left(\frac{1}{i} + \frac{1}{n-i+1} \right)^{-1/2} \left(\frac{S_i}{i} - \frac{S_n - S_{i+1}}{n-i+1} \right) \tag{3.6}$$

We propose a slightly adjusted statistic. Denote $\hat{\sigma}_i$ as the estimated standard error of the first i residuals, and $\hat{\sigma}_{i^-}$ as the standard error of the last i residuals. Let $Z^* = max_{1 \leq i \leq n} |Z_i^*|$ for

$$Z_i^* = \left(\hat{\sigma}_i^2 + \hat{\sigma}_{(n-i+1)^-}^2 \right)^{-1/2} \left(\frac{S_i - \frac{S_i}{i}}{i} - \frac{S_n - S_{i+1} - \frac{S_n - S_{i+1}}{n-i+1}}{n-i+1} \right) \tag{3.7}$$

Z^* greatly outperformed Z in both the simulations and observed data. We developed Z^* to account for the "messiness" in data normally encountered by political scientists. The residuals to the left and right of a given break point may be neither mean zero nor homoscedastic, and Z^* accommodates these possibilities while Z does not. This better captures the nature of observational data, but our test is no longer uniformly most powerful, due to the unbalancedness of the design and ambiguity of degrees of freedom in the variance estimate.

Since we are not as interested in the inference problem, we need only assume that the data are reasonably symmetric about the mean function. The statistic Z^* serves to locate the most likely position of the discontinuity. It considers, at each possible breakpoint, the extent to which the residual immediately to the left is above the mean to the left, and the residual immediately to the right is above the mean to the right. The more these two terms differ, the more likely a discontinuous break occurs at each point. The weight term in front

comes from the variance estimate, and experimentation revealed it effective in regulating erratic behavior at the boundaries.

This is a rough approximation. The pattern of residuals about the jump are not constant, but in the interest of finding a single break, binary segmentation is a reasonable, computationally simple approximation. As our examples below show, it is also quite feasible. We had experimented with modeling the strength of the break through a weighted average that aligns with residual patterns (e.g., wavelets, exponentially damped sine curves) as well as more exotic variable selection methods (Efron *et al.*, 2004b, e.g.), but, the additional complexity of method added little to the performance of the algorithm. When choosing jumps, the simple statistic above performed as well as its more complex competitors.

3.2.3 Stopping rules

Since this is an automated process, we require a stopping rule. We consider several here. The first is a modified BIC statistic, with $\hat{\sigma}_\epsilon^2$ an estimate of residual variance, n the sample size, n^* the number of randomly selected knots,[3] and k the number of jumps:

$$\arg\min_k \frac{\hat{\lambda} \cdot \int \{\hat{f}''(t)\}^2 dt}{\hat{\sigma}_\epsilon^2} + k \cdot \log(n) - \frac{k}{2} \cdot \log(n^*) + \frac{k}{2} \log(2\pi) \tag{3.8}$$

Appendix A contains a derivation of the traditional BIC statistic and our modification of it. Note how, in the event that the spline is fit to every observed data point ($n = n^*$), this statistic reduces to the traditional BIC statistic. The term with the $\log(2\pi)$ is asymptotically negligible, since it does not grow in n or n^*, but simulations revealed that the term was helpful in reducing the false positive rate for our smallest simulation ($n = 100, n^* = 30$). It did not make a large impact on the power of the larger simulations ($n = 250, 500$).

The first term is increasing in k through both the numerator and the denominator. As more jumps are added, the spline has to bend less in order to accommodate the jumps, so the spline fit is more linear, and the numerator increases. The penalty in the numerator is left unit-free by dividing through by the residual variance estimate. As well, the residual

[3]Splines select a random subset of knots due to computational difficulties in inverting large matrices. Even with relatively few knots, estimates are quite stable and accurate.

estimate decreases with the addition of more jumps, decreasing the denominator. The same logic motivating BIC then leads to a "cost" of $\log(n) - \frac{1}{2}\log(n^*)$ for the addition of each additional jump, in order to ensure convergence to the true model as $n \to \infty$. The spline penalty term serves the role of the log-likelihood, acting as a measure of the divergence between the "true" model and the fitted values.

A means for balancing model fit and size, the BIC provides an estimate to the Bayes factor. The BIC approximates the posterior probability that a given model is the true model, given a uniform prior over all candidate models. The statistic is not interpretable on its own, but is useful in model selection. The most preferred model with the sequential smoothing spline is the one with the smallest modified BIC statistic as given by equation 3.8.

A second stopping rule comes from the logic of hypothesis testing. Although we recommend our modified BIC statistic we present the following below. The reference distribution for the maximal Z is usually determined by permutation. This creates problems, though, because the reference distribution of each subsequent break must be conditioned on the selection of all previous breaks. The permutation method will increase in complexity rather quickly. As a simplification, we suggest considering the models as a sequential series of nested models. This suggests an approximate χ^2 statistic to consider among them:[4]

$$\frac{\hat{\lambda}\int\{\hat{f}''(t)\}^2 dt}{\hat{\sigma}_\epsilon^2} + c^*_{\alpha,k}$$

where c^* is the critical value of the χ^2 statistic, with family-wise error rate α and number of breaks k:

$$c^*_{\alpha,k} = \chi^2_{1-\alpha/k,k}$$

Using Bonferroni's adjustment of α, the term $1 - \alpha/k$ keeps the probability of making at least one type I error among all breaks below α. This approach has a noticeably lower

[4]Since we are not making any distributional assumptions about the error term, the χ^2 statistic is only an approximation.

threshold even at small sample sizes of 100 for the first several jumps; for this reason, we remain wary.

The final suggestion is using a combination of expert evaluation and common sense. As heuristics, we suggest a few options. First, if a selection may be a false positive, explore the selections afterwards. If several selections after it are known jumps, the jump under question is more plausible. If there are no known jumps afterwards, then its selection is less plausible. If the researcher has a specific prior distribution proportional to weights w_i over all possible jumps, the statistic $w_i|Z_i^*|$ may of course be used instead. Similarly, known breaks can be incorporated into the unpenalized space directly, and our method consequently implemented.

Second, we noticed that sometimes the BIC may be quite flat around the minimum. This may limit the discussion to the range of acceptable jumps, such as 3-5, even if the strict minimum occurs at 4. Similarly, the first few jumps result in a large decrease in the penalty term, but the effect decreases dramatically. Using either rule above, some jumps are clearly reasonable, some are questionable, and some are unreasonable. We suggest either reporting the BIC statistic or χ^2 p-values along with the data, and the researcher may decide to take the first of the "questionable" jumps as the stopping points. If computational power permits, resampling methods may be used, but we have been satisfied with the performance of choosing the modified BIC-minimizing number of partial splines. That is the rule we use throughout the remainder of the paper.

Finally, with reasonably large datasets, the cubic splines are fit only using a random subset of the data as "knots." This introduces variance into the stopping rule statistic, so we recommend exploring either adding to the number of data points selected or a resampling method to help get a sense of the variance of the test statistics described here. With large datasets, of over one thousand, increasing the number of knots can aid in detection, making the acceptance level of our BIC statistic less stringent. Increasing the number of knots, though, is not computationally prohibitive. Bush's approval data discussed below has an n of 1533, and `ssanova` selects only 52 knots. Computation took about 12 seconds on a

Pentium D processor. Using our algorithm with 200 subsampled datapoints lengthened the time to only two minutes and eleven seconds.

These guidelines may sound arbitrary, but, as illustrated below, our method is never "too" wrong. It is conservative, in that it rarely over-estimates the number of jumps, and even with reasonably small sample sizes and in the presence of auto-correlated noise, the method is powerful.

3.2.4 Procedure statement

We suggest the following recursive procedure for finding these jumps in smooth functions:

1. Fit a cubic smoothing spline using the REML algorithm to estimate the smoothing parameter.
2. Remove the linear trend from the residuals.
3. Search the residuals for a breakpoint, using the binary segmentation algorithm.
4. Add the corresponding break to the partial spline's unpenalized space.
5. Repeat from 1 until the BIC stopping rule ends the procedure.

This will result in breaks $\delta_1 \dots \delta_k$, and a mean function estimate of the form:

$$\hat{y}_i = \hat{f}(x_i) + \sum_{j=1}^{k} \hat{\beta}_j I(x_i > \delta_j)$$

where $\hat{f}$ is a smooth spline estimate. The estimate $\hat{\beta}_j$ gives the estimated magnitude of the discontinuity, and is reported when fitting partial splines. The standard errors on this coefficient will be wrong, though, since they do not account for the sequential nature of the selection and fitting. Were this magnitude of interest, and confidence intervals desired, we recommend a parametric bootstrap. Our algorithm sequentially augments the unpenalized space of the spline, fitting fixed jumps at specific points. Although estimating the number of jumps simultaneously may be preferred, and in more than one dimension is most certainly

preferable, we are happy with the performance of the method in one dimension. We illustrate the procedure below on two observed datasets, Bush's approval and DW-NOMINATE scores, as well as through a series of simulations.

3.3 Case 1: George W. Bush's Approval

The data in this section are 1533 estimates of George W. Bush's approval, from 33 different houses, between January 25, 2001 to October 24, 2007. We observe polls on 1041 dates over the 2463 dates total. Although the polls are taken over multiple days, we consider the date range to be the end date on which the poll is conducted. The approval estimates range from 27 to 89 percent. The data are all publicly available.

Figure 3.3 shows the fit to each of the different methods, each in R. We use a moving average with a two-week window, loess with spans .1 and .01 with command `loess` in library `stats`, the smoothing splines from command `ssanova` in library `gss` , a structural time-series model using a Kalman filter from command `StructTS` in library `stats`, and our sequential segmentation method.

Results from each of the methods are shown in figure 3.3. Our sequential segmentation method outperforms, in terms of mean squared error (MSE), both the loess with span .1 and the smoothing splines. Our method performs worse, in terms of MSE, than the Kalman filter and the loess with span .01, but the improvement in MSE comes at the cost of a rather jagged fit. Figure 3.4 shows a larger graph, comparing splines to the sequential segmentation method. Notice how splines smooth over non-"smooth" events. For example, the bump in Bush's ratings after 9/11 could not have started before the event; people had no expectations

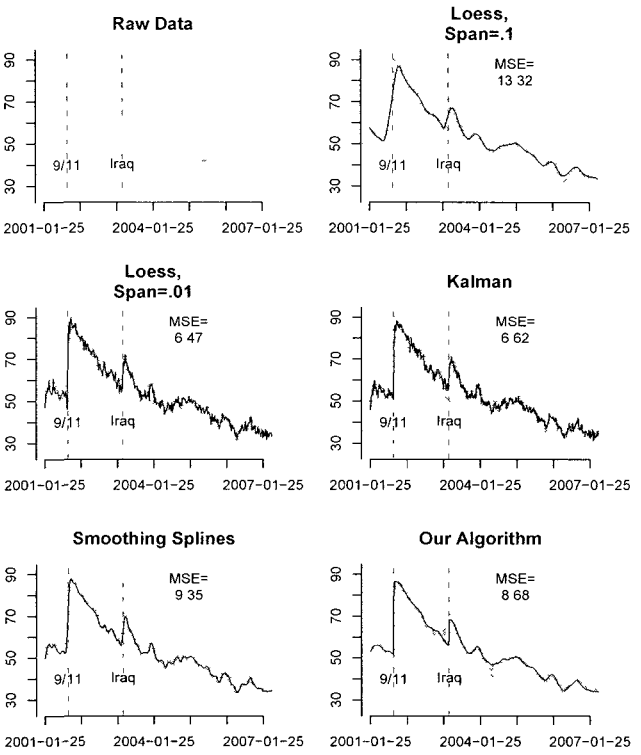

Figure 3.3: A comparison of fits to the data among the different smoothing methods.

that the events would occur on a certain date and so they did not start revising their opinions of Bush upwards before then. Note also how the "rally around the flag" effect, whereby a President's popularity increases upon starting a war, is immediate. During the run-up to the war, Bush's approval dropped regularly, but immediately upon the invasion, he got a bump up that then proceeded to erode. Our algorithm selected the "rally around the flag" effect as instantaneous, even when the invasion of Iraq was clearly expected.

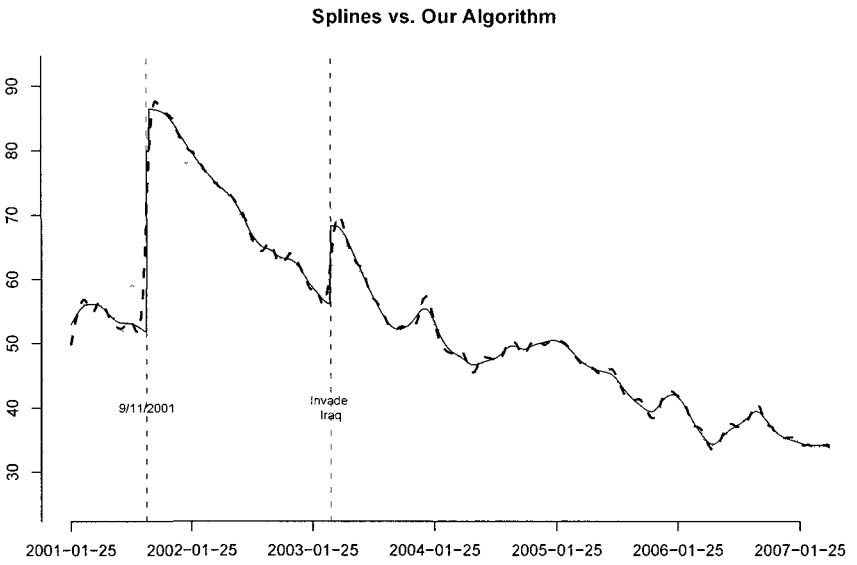

Figure 3.4: Spline fit is dashed; our method fit is solid.

Just as important, note how our method follows the spline method in areas after the events. From about mid-2004 on, our method follows the smoothing spline almost exactly. Our sequential segmentation method acts like a smoothing spline when appropriate.

The real payoff for the sequential segmentation method comes from looking at autocorrelation and the QQ plots. Figure 3.5 shows the QQ plots, compared to a normal distribution, for the residuals using each of the six methods earlier. The other methods tend to do passably well in either having normally distributed errors (the second loess fit) or in having

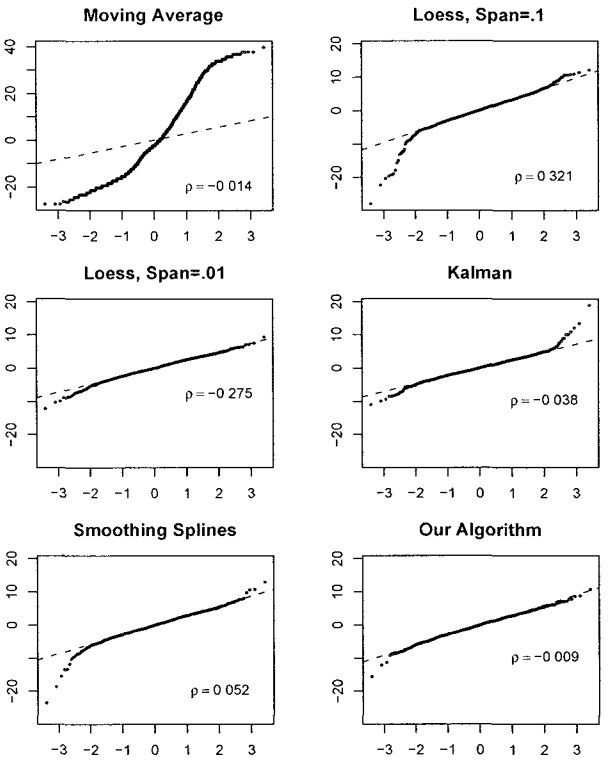

Figure 3.5: QQ plots, for a normal distribution, for each of the smoothing methods. "ρ" gives the correlation in the residuals with lag 1. For ease of interpretation, the last five plots are on the same y axis.

low auto-correlation (the moving average, Kalman filter, and splines). Only our algorithm does well on both counts, and noticeably better than the other methods.

Next, we consider the events selected as "jumps" in approval. The jumps and their corresponding BIC statistics are listed in table (3.1). The trouble with evaluating whether these jumps are valid is twofold. First, we have no idea what the "true" curve is, or even if the concept of such a curve is theoretically coherent. We conceive of this curve, instead, as

Selected Date	Event	Actual Date	BIC
Spline Only	XX	XX	78.05
2001-09-12	9/11	2001-09-11	68.62
2003-03-20	Invade Iraq	2003-03-20	60.78
2001-09-21	Bush's Post 9/11 Address to Congress	2001-09-21	63.67
2004-01-22	Bush's State of the Union	2004-01-22	72.17
2006-10-04	??	??	78.12

Table 3.1: The first five events in Bush's term selected by the sequential segmentation spline. The BIC criterion selects the first two events. We include the next three events to illustrate behavior of the BIC statistic.

the consensus of approval ratings among pollsters. Looking at the dates chosen and seeing if anything happened involving approval considerations of Bush leads to a problem, though: given any day, you can find *something* of import that might affect approval towards Bush. Unfortunately, there is no objective listing, serial or otherwise, to which we can compare these results.

That said, we are confident in the dates selected. We are counting the date of each poll as the last day it was administered, so we look for critical events a few days before the breaks selected by our method. The two dates selected, 9-12-2001 and 3-20-2003, clearly correspond with major events that would plausibly affect approval scores in the Bush presidency (9/11, the invasion of Iraq on 3/17). Since we are taking the date of the poll as the final day of its administration, the method selects dates a few days past the event of interest.

The primary substantive conclusion is that presidential approval, and quite likely other public mood trends, are not smooth. While recent theoretical work on macropolitical evolution views Presidential approval as a smooth, AR(1) time series in Gallup approval (Erikson

et al., 2002), we find that future empirical work should search for and model multiple structural breaks.

3.4 Case 2: Congressional Ideology

The data in this section are three time series: the median estimates of the first dimension of DW-NOMINATE ideology scores by Congress and party. We consider two parties, the Democrats and the non-Democrats, a pooling of Republicans and Whigs for the sake of this paper. We have data on each Congress from 1788 to 2004,[5] the non-Democratic Party from 1800-2004, and the Democratic Party from 1794-2004. Higher scores are commonly interpreted as being more "conservative" on economic issues, while a lower scores means that the Congressional median is more "liberal" on economic issues (Poole and Rosenthal, 1997). The data are publicly available.

Unlike Bush's approval data, the median DW-NOMINATE data is evenly spaced through time, although with only one data point per Congress. Relative to Bush's approval, this is a relatively small-n test of the method, with only 109 data points, versus 1533 above.

A second key difference is that, in this data, no apparent discontinuities appear. The raw data for the entire Congress can be found in the upper left hand corner of figure 3.6, with fits from various methods in the subsequent boxes. We notice nothing in this data comparable to the "9-11" problem of above.

Despite the lack of visually obvious jumps, a long literature has debated as to whether some elections may be "critical" (Key, 1955, 1959; Mayhew, 2002). Most Congressional

[5]For clarity, we refer to each Congress by the year in which it was elected rather than its number, i.e. "2006" rather than "110th."

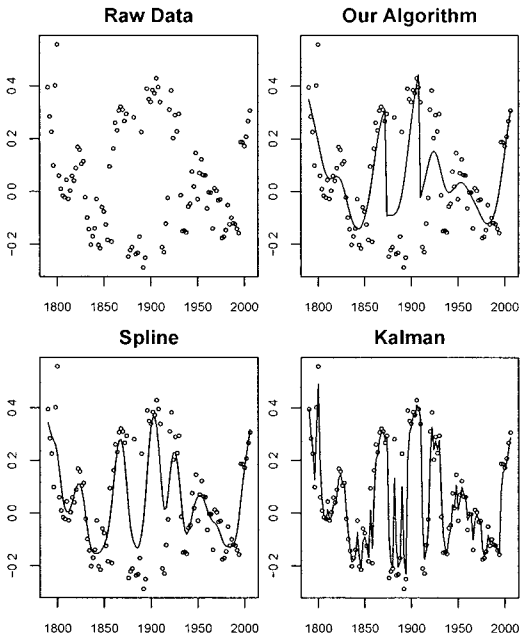

Figure 3.6: A comparison of fits to the median DW-NOMINATE score by Congress among three different smoothing methods.

elections are "secular," reflecting slow gradual change, while some scholars consider a select few as "critical," in that they result in rapid, sustained shifts in partisan composition of the public. This dynamic, as expressed through Congress members, captures the nature of the "smooth+jump" type function that we are estimating here.

As suggested by the realignment literature, we rely rather heavily on our model's assumptions: that median NOMINATE score is the sum of a smooth and discontinuous part.

	Date	Event	BIC
	None	Spline Only	23.79
Entire Congress	1910-12*	Joseph Cannon Rebellion	22.31
	1874-76*	End of Reconstruction	21.58
	1932-34	New Deal Realignment	23.11
	1896-98	1896 Realignment	26.87
	1860-62	Civil War	29.74
	Date	Event	BIC
	None	Spline Only	20.90
Non-Democratic Party	1818-20*	Possible False Positive?	19.67
	1852-54	??	22.79
	1856-58	??	25.05
	Date	Event	BIC
	None	Spline Only	14.92
Democratic Party	1930-32*	New Deal Realignment	14.59
	1892-94	Populist Realignment	17.53
	1954-56	??	19.87

Table 3.2: The first few events in Congressional history selected by the sequential segmentation spline. Dates marked with an asterisk are selected by our BIC criterion.

Comparing the fit of our method to the spline and time series method highlights how statistical findings are dependent on the assumptions under the statistical model. A researcher searching for smooth cycles in Congressional history may prefer either of the bottom two graphs. A researcher searching for a smooth curve with a few jumps would prefer our model. Our algorithm, though, is both consistent with the data, and it reduces to a spline in the limiting case.

Most of the discovered jumps, as shown in table (3.2), are easily attributable to commonly accepted shifts in Congressional behavior. The two shifts selected for the entire Congress via BIC are 1910-1912 and 1874-1876. The first can be attributed to the rebellion against Speaker Joseph Cannon, which shifted power from the Speaker to the committee chairs.

The second can be attributed to the end of Reconstruction. The next few dates, 1932, 1896, and 1860, though not selected by BIC, have long been considered the canonical dates for "realignment" (Mayhew, 2002). The 1930-1932 break for Democratic median votes is clearly attributable to the New Deal. A single shift, in 1818-20, was selected for the non-Democratic party. We worry this may be a false positive, and may be a random artifact of pooling all non-Democratic parties.

The algorithm performed well, even in this noisy, imprecise setting. Plotting the data revealed no apparent, obvious jumps, and condensing the NOMINATE scores down to 109 points is a gross over-simplification of Congressional behavior and evolution. Given all this, our method yielded reasonable results, selecting only one false positive and a series of dates otherwise that correspond with well known shifts in Congressional makeup and behavior.

3.5 Simulations

We conduct twenty-four separate simulations in order evaluate our method. We assume two different systematic components, where f is a Bessel function of the second type, and x in the interval $[0, 1000]$. We ran each simulation one thousand times each, comparing our algorithm to both smoothing splines (function `ssanova` in R library `gss` and a Kalman filter (function `StructTS` in R library `stats`. The characteristics of the simulations are below:

Model Specifications:

$$sim_{jump}(x_i) = -2 \cdot f(x_i/100) + 8 \cdot I(x_i > 200) - 4 \cdot I(x_i > 500) + 2 \cdot I(x_i > 800) + u_i$$

$$sim_{nojump}(x_i) = -2 \cdot f(x_i/100) + u_i$$

Variance Specifications:

Gaussian Noise: $var(u_i) \in \{1, 4\}; cor(u_i, u_{i-1}) = 0$

AR(1) Noise: $var(u_i) \in \{1, 4\}; cor(u_i, u_{i-1}) = .4$

Sample Sizes Used:

$$n \in \{100, 200, 500\}$$

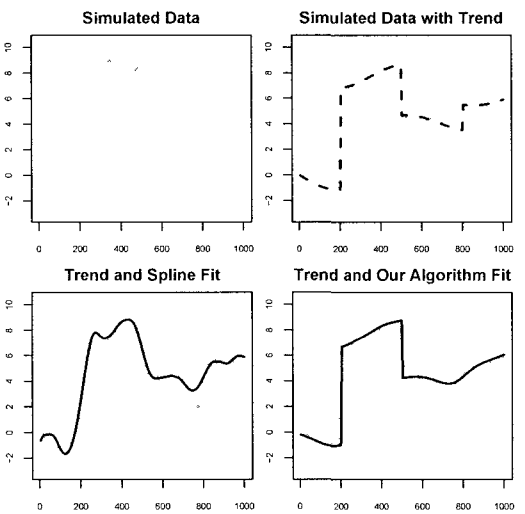

Figure 3.7: An example from a run of the simulation, with n=200, Gaussian noise, and var(u_i)=1.

Figure (3.7) contains an example of the simulation with jumps as well as one of our fits to it. The first jump is significant at the five percent level in all situations. The second jump is significant only in our less noisy simulations, and the third jump is never significant at five percent.

Simulations With Jumps									
	n	var (u_i)	0	1	2	3	4	5	6
Gaussian	100	1	63.2	27.2	6.6	2.4	0.6	0.0	0.0
	200	1	7.0	56.8	36.2	0.0	0.0	0.0	0.0
	500	1	0.8	7.5	89.2	2.5	0.0	0.0	0.0
	100	4	68.9	18.6	12.5	0.0	0.0	0.0	0.0
	200	4	36.3	39.2	24.5	0.0	0.0	0.0	0.0
	500	4	7.2	74.5	18.3	0.0	0.0	0.0	0.0
AR(1)	100	1	67.6	23.2	6.5	2.1	0.6	0.0	0.0
	200	1	19.6	32.7	39.6	7.3	0.8	0.0	0.0
	500	1	8.3	5.5	45.4	39.0	1.8	0.0	0.0
	100	4	34.1	41.9	20.6	2.9	0.5	0.0	0.0
	200	4	22.8	48.8	25.5	2.5	0.4	0.0	0.0
	500	4	10.3	33.3	51.1	4.7	0.6	0.0	0.0
Simulations Without Jumps									
	n	var (u_i)	0	1	2	3	4	5	6
Gaussian	100	1	89.2	10.8	0.0	0.0	0.0	0.0	0.0
	200	1	88.9	8.8	1.3	0.6	0.3	0.1	0.0
	500	1	99.4	0.2	0.1	0.0	0.3	0.0	0.0
	100	4	100.0	0.0	0.0	0.0	0.0	0.0	0.0
	200	4	96.9	1.8	0.8	0.2	0.3	0.0	0.0
	500	4	99.9	0.1	0.0	0.0	0.0	0.0	0.0
AR(1)	100	1	92.8	6.5	0.7	0.0	0.0	0.0	0.0
	200	1	98.2	1.6	0.2	0.0	0.0	0.0	0.0
	500	1	99.7	0.3	0.0	0.0	0.0	0.0	0.0
	100	4	91.6	7.8	0.5	0.1	0.0	0.0	0.0
	200	4	92.3	7.4	0.3	0.0	0.0	0.0	0.0
	500	4	96.8	3.2	0.0	0.0	0.0	0.0	0.0

Table 3.3: Percent of times each number of breaks was chosen, by simulation. There are three jumps total, two of which are easily discernible and one which is not.

We vary the simulations by sample size, error variance, and variance structure. Table (3.3) contains a list of each of the simulations we conduct for both sim_{jump} and sim_{nojump}, as well as the percent of times each number of jumps was selected by our algorithm. The simulations all contain either independent, Gaussian noise or AR(1) noise, with correlation .4.

Our simulations demonstrate four desirable aspects of our algorithm. First, as shown in the bottom half of table (3.3), the algorithm has a low false positive rate when there are no jumps in the model. When there are no jumps, the algorithm will reduce to the spline, with an acceptable false positive rate. For small sample sizes, and in our less noisy scenarios, the false positive rate is approximately ten percent, but the false positive rate drops dramatically as the sample size gets to 500. False positives rates are acceptably low in the other simulations, even for $n = 100$. The method is robust against error misspecification, and the false positive rate drops to zero as the sample size increases. When there are no jumps in the true function, our method uncovers false positives at an acceptable rate.

Second, the method is powerful. Table (3.4) shows the percent of the time each of the three breaks were selected. The largest break, at $x_i = 200$, is selected the most. By $n = 500$, our algorithm selects the first break between 89%-99% of the time. The next two breaks are selected less than the first break, as expected, since they are smaller in magnitude. The discovery rate of each of the smaller breaks, though, is clearly increasing in n. All of these results are robust to error-misspecification. Just as importantly, the false positive rate is

	n	var (u_t)	ID 200	ID 500	ID 800	False Positive Rate
	100	1	84.4	14.8	0.0	0.6
	200	1	93.0	36.2	0.0	0.0
	500	1	99.2	91.6	2.4	0.2
Gaussian	100	4	30.7	11.4	0.0	1.5
	200	4	63.6	22.1	0.0	2.5
	500	4	92.8	17.8	0.0	0.5
	100	1	34.0	7.9	1.1	1.9
	200	1	83.0	47.7	5.8	0.5
	500	1	93.5	86.2	40.1	0.7
AR(1)	100	4	67.3	20.4	0.8	5.3
	200	4	77.4	27.3	0.9	3.3
	500	4	89.8	55.8	4.1	2.3

Table 3.4: Percent of the time each break was identified, and the false positive rate, by simulation.

acceptable, always below 5.5 and well below 1% in five of the twelve specifications with jumps.

Third, the algorithm provides, on average, gains in squared error, and at best, the algorithm results in only modest losses versus the smoothing splines and time series method, as illustrated in table (3.5). We calculated squared error as the sum of the squared difference between the fitted and true values. As shown in the top half of table (3.5), our method performs well relative to the other methods. The Kalman filter proves marginally better, about 4%, in two of the twelve simulations with jumps. Improvements in squared error over the Kalman filter and splines in the remaining simulations range between 1% and 55%. Of the twelve simulations with jumps, our algorithm outperforms the Kalman filter by more than 20% nine of the twelve times. Our algorithm similarly outperforms the splines by more than 20% six of the twelve times.

Simulations With Jumps							
	n	var (u_i)	MSE, Kalman	MSE, Spline	MSE, Proposed	% Improved, vs. Kalman	% Improved, vs. Spline
	100	1	0.609	0.696	0.632	−3.83	9.20
	200	1	0.481	0.513	0.380	20.89	25.86
	500	1	0.338	0.355	0.158	53.16	55.42
Gaussian	100	4	1.549	1.234	1.192	23.05	3.44
	200	4	1.156	0.879	0.718	37.85	18.28
	500	4	0.763	0.566	0.347	54.59	38.73
	100	1	3.641	3.648	3.601	1.10	1.30
	200	1	0.356	0.453	0.368	−3.38	18.85
	500	1	0.306	0.330	0.166	45.85	49.79
AR(1)	100	4	1.316	1.036	0.937	28.82	9.57
	200	4	1.138	0.753	0.590	48.20	21.74
	500	4	0.942	0.506	0.287	69.56	43.27
Simulations Without Jumps							
	n	var (u_i)	MSE, Kalman	MSE, Spline	MSE, Proposed	% Improved, vs. Kalman	% Improved, vs. Spline
	100	1	0.168	0.069	0.082	51.31	−18.43
	200	1	3.631	3.434	3.405	6.21	0.84
	500	1	3.527	2.978	2.970	15.80	0.27
Gaussian	100	4	0.444	0.260	0.261	41.36	−0.05
	200	4	13.867	12.369	12.336	11.04	0.27
	500	4	13.897	11.080	11.084	20.24	−0.04
	100	1	0.215	0.076	0.080	62.70	−5.62
	200	1	0.174	0.044	0.045	74.17	−2.44
	500	1	0.137	0.019	0.019	86.03	0.18
AR(1)	100	4	0.723	0.265	0.273	62.31	−3.01
	200	4	0.600	0.145	0.151	74.81	−4.21
	500	4	0.494	0.065	0.067	86.38	−3.45

Table 3.5: Mean squared error across simulations. The last two columns show the average percent improvement of our algorithm over the other two methods.

The bottom half of table (3.5) illustrates the results when the systematic component does not contain jumps. In every instance, our algorithm outperforms the Kalman filter, ranging between 6%-83% gains in squared error loss. Our method performs comparably to the smoothing spline. The only exception is in the smallest, least noisy simulation, where

splines outperform our algorithm by 19%. Of the remaining eleven simulations, the spline never provides an improvement of more than 6%, and in six of the twelve simulations, our algorithm differs from the smoothing spline by less than 1%.

Finally, the method is robust with respect to correlated errors. While our algorithm is less powerful in the presence of AR(1) noise around the discontinuous, smooth function, by modest sample sizes of 200, the algorithm performs well. Even with AR(1) noise, our method maintains a small false positive rate. In the presence of breaks and AR(1) noise, our algorithm outperforms splines in each case, and in the absence of breaks, splines never provide more than a 6% advantage in squared error over our algorithm. In two of the twelve simulations with AR(1) noise, our algorithm performs comparably to the Kalman filter; in the remaining ten, it provides substantial improvement, ranging between 25%-86%.

3.6 Extensions of Method

Our method admits several interesting and useful extensions. First, we have presented it so far as a modeling, rather than inferential, tool, although resampling methods could provide confidence intervals. Since the search for breaks is sequential, the nature of the distribution under the null hypothesis is not clear; each subsequent break is conditioned on the occurrence of the earlier breaks. Entering unpenalized covariates of interest into the nonpenalized space also allow for inference in a semi-parametric setting.

Second, our method can extend to a broad array of spline models.[6] These include extensions to higher dimensions, through thin-plate splines (Pearce and Wand, 2006). Our

[6]Specifically, any function that lives in a Hilbert space and has a reproducing kernel. See Wahba (1990) for extensions.

method can help find jumps in geographic data, or any situation in which the context dictates estimating a functional form that is smooth with breaks. Since the spline is a limiting case, and our cases above show that the method can "act like a spline" when it needs to, we hope to generalize the algorithm to where it can be useful across a broad array of data sets and questions.

Third, we have so far characterized our jump covariates as a series of indicator functions. The result is a form of a sequential cumulative sum tests, with a spline fit in between each identification of a jump. The jumps, though, could instead contain covariates that are of interest to the researcher. For example, using our Bush data, we could look for jumps in approval as a function of economic news, Congressional approval, etc. This would require simply ordering the outcome by a different covariate than time and using our sequential segmentation spline along this dimension.

Finally, we have developed the method so far with few constraining distributional assumptions about either the systematic component or the error. Stronger assumptions can be made, as necessary, in order to characterize a likelihood that could handle choice models, count models, and other limited dependent variable models.

3.7 Conclusion

This project began with the goal of estimating house-level effects on presidential approval. In estimating these effects, it quickly became apparent that extant smoothing methods missed the underlying dynamic of presidential approval. In fact, within much political data of interest, "smoothing" appears to be presumptuous. Many processes are a mixture

of both slow-moving change and immediate jumps due to rapid shocks. Smoothers miss the nature of the underlying function, while structural break tests offer no stopping criterion. This led us to consider first, how to think about these processes, and, second, how to find the breaks and when to stop doing so.

Our method has performed well in both simulations and on observed data. Within a dense presidential approval data set, our method discovered breaks that correspond with clearly identifiable shocks. Within a relatively sparse Congressional data set, it was equally successful at highlighting important dates. The simulations further reenforce our faith in the method.

The "smooth+jump" function we describe here could apply to a wide variety of political processes that face critical events. The algorithm allows a flexible means to model other social and physical processes of interest. Examples include change-point models, or any model that must accommodate some smooth curve with the occasional persistent shock. In the face of a known exogenous shock, but uncertainty over the particular timing of the impact of the shock, our sequential segmentation spline will allow for estimation of the existence and most likely location of the discontinuity.

Testing for the existence of a structural break at a given point is straightforward. Here, we provide a method for solving a far harder problem: searching through all possible breaks, adding jumps sequentially, and stopping at a reasonable point. We plan to extend the method in the future to limited dependent variables and higher-dimensional settings.

Chapter 4

Identifying the Effects of Political Boundaries

4.1 Introduction

Political scientists have made increasing use of geographic data to identify jurisdiction-specific effects (Berry and Baybeck, 2005; Keele and Titiunik, 2011; Ward and O'Loughlin, 2002). Identifying these effects poses two problems. The first problem is well-understood: the correlation among geographically proximate units must be modeled (Bivand *et al.*, 2008). This requires methods that capture similarities among nearby units. The second problem is that models with a large number of jurisdiction-specific intercepts and slopes can grow unwieldy. The proposed method addresses both concerns. Geographic correlation is modeled with a smoother, while the most important jurisdiction-specific effects are *selected.* This allows researchers to consider models with hundreds of covariates but with most of their effects estimated at zero. Normally unmodeled subtleties in the data can be uncovered in a statistically rigorous manner.

Identifying these effects requires accounting for local correlation (see Beck *et al.* (2006) for an overview), so as not to confound jurisdiction-specific effects with regional effects. To illustrate, consider the map of 2008 US Presidential electoral outcomes in figure 4.1. There

are regions both "red" (Republican) and "blue" (Democratic). States won by the Democrat Barack Obama clustered on the coasts, Midwest, and the Sun Belt , while the Republican candidate, John McCain, captured states through the Great Plains and South (figure 4.1, top). Looking at county-level data reveals a more subtle phenomenon (figure 4.1, bottom). Urban, coastal, and strips of the upper Midwest appear solidly Democratic, while Appalachia and portions of the South and Rocky Mountain states appear solidly Republican with no obvious discontinuities at state lines. The two maps in figure 4.1 highlight the question: do regional partisan differences respect state lines?

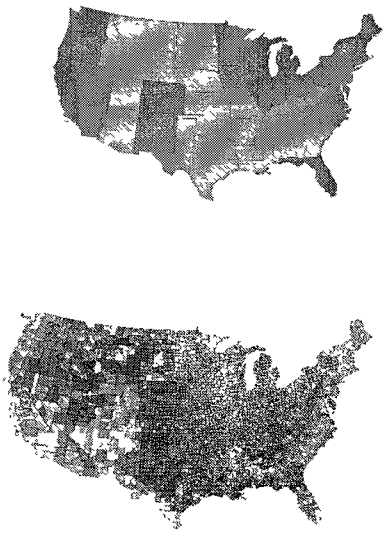

Figure 4.1: State-level and county-level returns from the 2008 Presidential election. Darker colors correspond with areas relatively supportive of Barack Obama; lighter colors denote those areas relatively more supportive of John McCain. Areas and colors were not adjusted for population size.

The proposed method is most applicable when the researcher suspects systematic heterogeneity across jurisdictions but has little a priori theoretical guidance as to which states indeed have an effect. Rather than simultaneously estimate a large number of coefficients,

the proposed method selects a subset of relevant effects. Existing hierarchical methods that fit these models, as in Gelman and Hill (2006), are difficult to both interpret and present because too many coefficients are returned. For example, a model with a state-specific effect and two state-specific covariates for the continental United States returns 147 coefficients ($=49 \times 3$). Current best practice for interpreting and presenting these coefficients, as exemplified in Gelman *et al.* (2008), involves a series of three different plots, where each plot contains a 7×7 grid of the outcome versus the state specific effect. The proposed method remedies this by producing coefficients for only a subset of a (possibly vast) number of variables, producing a clearer picture of what should be interpreted. Setting most of the coefficients to zero serves to select relevant variables, and their coefficients can be interpreted in a normal manner.

The variable selection method employed is the *L*east *A*bsolute *S*hrinkage and *S*election *O*perator, or LASSO (Tibshirani, 1996). The LASSO produces point estimates of zero for most coefficients, thereby selecting the most relevant effects. The LASSO is a penalized regression method, whereby a linear model is fit subject to a constraint on the sum of the absolute values of the parameters. As discussed below, this has the desirable property of producing point estimates of *precisely* zero for most effects. In practice, the method and its extensions have been shown to be a powerful means of identifying a meaningful subset of variables. The LASSO has generated a vast literature, across fields from statistics and computer science to biology and public policy (Hesterberg *et al.*, 2008). Political scientists

have been remarkably silent in this field; this paper introduces political scientists to many of these insights.

The geographic correlation is modeled through the popular nonparametric method of smoothing splines (Wahba, 1990; Gu, 2002; Shawe-Taylor and Cristianini, 2004). Smoothing splines work through specifying a set of smooth covariates (basis functions), and then fitting the data to these smooth bases. A parameter is introduced that controls the level of "curviness"' of the resultant fit, and this parameter is selected to balance model fit with model complexity. The proposed method integrates variable selection and smoothing, allowing the researcher to fit a broad class of models while returning parsimonious, interpretable results.

The chapter progresses in five parts. First, the methodology is described, illustrating how smoothing splines and variable selection can be combined into a single optimization problem. Second, the algorithm and fit criterion are then described. Third, a set of simulations illustrate the method's efficacy. Fourth, the proposed method is then applied to two applied datasets: partisan outcomes in the 2008 Presidential election and economic output in Africa. Fifth, a conclusion follows.

4.2 Smoothing and Variable Selection Methods

The proposed method works through combining two different types of penalized regression. The first component penalizes the sums of the squares of the spline coefficients, which naturally "smooths" the resultant curve. The second penalizes the sums of the absolute values of the jurisdiction-specific coefficients, which forces many coefficients to have point estimates of precisely zero. This section provides a brief overview of the two methods. The

following section contains the proposed method, a combination of smoothing and variable selection.

4.2.1 The Smooth Component

A researcher often does not know a reasonable functional form for her target function. She may know it to be some function of an observed variable, but she may know little about whether the effect is linear, quadratic, cubic, and so on, in the observed data. To handle this uncertainty, I model the smooth geographic component through the use of the popular nonparametric method of smoothing splines (Wahba, 1990; Gu, 2002; Shawe-Taylor and Cristianini, 2004; Scholkopf and Smola, 2001; Pearce and Wand, 2006). Applications of this approach aimed explicitly at political scientists can be found in Keele (2006, 2008) and Beck and Jackman (1997).

The simplest smoothing spline models fit a model additive in a linear and a nonlinear component. The linear component is parametric. It is assumed to be linear in some small set of variables. The nonlinear component is nonparametric, as it can handle a large class of arbitrarily smooth curves. A parameter is introduced to guard against overfitting, balancing the tradeoff between a too-complex model that overfits the data and a too-simple model that misses a systematic trend.

More formally, assume a vector of n outcomes, y. Assume an $n \times k$ matrix, S, that models the parametric component. In the cases illustrated in this chapter, columns of S contain the latitude and longitude of the observation, an intercept, and any controls. Denote the vectors of latitude and longitude as s_{lat} and s_{long}, respectively, with each vector rescaled to

fall in the unit interval. These vectors are assumed to be a realization from random variables S_{lat} and S_{long} distributed uniformly and independently over the unit square. Assume as well a smooth, continuous function of the latitude and longitude, $\eta(s_{lat}, s_{long})$. This function captures the geographic correlation among the observations, by expressing the outcome, y as a function of its proximity to other observations. Finally, for simplicity, assume a vector of symmetric, independent, and equivariant disturbances ϵ. The outcome, y is then modeled as

$$y = Sd + \eta(s_{lat}, s_{long}) + \epsilon \tag{4.1}$$

A measure of the complexity, or "curviness" of η, denoted $\Omega_{tps}(\eta)$ is introduced. The subscript *tps* denotes the "thin plate spline," the most common two—dimensional model. The complexity measure characterizes the total curvature of the fitted curve over its domain.

Before discussing the two dimensional case, consider the simpler one-dimensional case of the cubic smoothing spline. In this case, $\eta(t)$ is written as a function of some variable t, where t has support over $\mathcal{T}$. For the cubic spline, the complexity measure $\Omega_{cubic}(\eta)$ is given as

$$\Omega_{cubic}(\eta) = \int_{\mathcal{T}} (\eta''(t))^2 dt \tag{4.2}$$

The two-dimensional thin plate spline is a natural extension of the cubic spline. Because the proposed method fits a curve over two dimensions, latitude and longitude, the second derivatives and the cross derivative are squared and integrated over their support. This gives

a measure of the form

$$\Omega_{tps}(\eta) = \int_{[0,1]} \int_{[0,1]} \left(\frac{\partial^2 \eta}{\partial^2 S_{lat}} \right)^2 + 2 \cdot \left(\frac{\partial^2 \eta}{\partial S_{lat} \partial S_{long}} \right)^2 + \left(\frac{\partial^2 \eta}{\partial^2 S_{long}} \right)^2 dS_{lat} dS_{long} \qquad (4.3)$$

The integrand can be seen to be everywhere nonnegative, as it is the sum of three squared terms, and each summand is the square of η differentiated twice. The integral sums the magnitude of squared second derivatives ("curviness") at each value of S_{lat} and S_{long}. The objective function can now be written as the form

$$\underset{d,\eta}{\text{argmin}} (y - Sd - \eta(s_{lat}, s_{long}))'(y - Sd - \eta(s_{lat}, s_{long})) + \lambda_2 \Omega_{tps}(\eta) \qquad (4.4)$$

The tuning parameter, λ_2, controls the level of smoothing. Selecting $\lambda_2 = 0$ produces a too-complex model that completely interpolates of the data, i.e. $y = \hat{y}$. As $\lambda_2 \to \infty$, the fitted values approach the least squares line from regressing $\mathbf{y}$ onto S. Selecting λ_2 controls the balance between these two extremes.

The celebrated Representor Theorem of Kimeldorf and Wahba (1971) shows that the population minimizer of the form $E((y_i - \hat{y}_i)^2 | S)$ can be written as $\hat{y} = R\hat{c} + S\hat{d}$. R is a"reproducing kernel," an $n \times n$ matrix purely determined by s_{lat} and s_{long} and assumptions about the nature of η. R is selected to serve two goals simultaneously. First, it provides a means to model the geographic correlation; each element of R is a measure of proximity between the two points. Second, R is selected to provide a set of basis functions to parameterize the smooth curve.[1] The most common two-dimensional reproducing kernel is that of

[1]More technically, assume η lives in the Hilbert space of all smooth curves defined over the unit square, $\mathcal{H}$. Hilbert spaces are compact, so they contain all their limit points, and have an inner product, which gives a sense of angle. Next, assume $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$. $\mathcal{H}_0$ is the finite dimensional space spanned by S. Given the

the thin-plate spline. The reproducing kernel is calculated as

$$R = [r_{ij}] = \left\{(s_{long,i} - s_{long,j})^2 + (s_{lat,i} - s_{lat,j})^2\right\} \log \left\{(s_{long,i} - s_{long,j})^2 + (s_{lat,i} - s_{lat,j})^2\right\} \tag{4.5}$$

R is the penalized component, parameterizing the smooth curve, while S is the unpenalized component. With known R and S, the problem reduces to a problem of the following form (Wahba, 1990):

$$\{\hat{c}_{SS}, \hat{d}_{SS}\} = argmin_{c,d}\ (y - Rc - Sd)'(y - Rc - Sd) + \lambda_2 c' Rc \tag{4.6}$$

Because R is an $n \times n$ matrix, the problem has more parameters $(n+2)$ than observations (n), necessitating the constraint. λ_2 balances the trade-off between the least squares fit and a complete interpolation of the data. As shown below, variable selection also requires a similar trade-off between model fit and model parsimony.

4.2.2 The Variable Selection Component

Every researcher has had to address the question of which variables to include in a model. A host of questions are nearly always left unanswered: why no interactions, or only the small number provided? Why no quadratic or cubic terms? The variable selection literature strives to provide a rigorous, and data-driven, answer to this questionHastie *et al.* (2001a). I focus primarily on variable selection through the Least Absolute Selection and Shrinkage Operator, or LASSO (Tibshirani, 1996).

---

eigenvalues of $\mathcal{H}_1$, $\{\gamma_i\}$, assume that these that satisfy $\sum_{i=1}^{\infty} \gamma_i^2 < \infty$. Denote D as the differential operator in $\mathcal{H}_1$, and its inner produce as $< \cdot, \cdot >_{\mathcal{H}_1}$. R is a symmetric matrix composed of eigenfunctionals of the twice-iterated Laplacian, D^2, evaluated at the data points. It is a particular Green's Function, evaluated on the data. This guarantees that $D^2 R = R$, $xRy =< x, y >_{\mathcal{H}_1}$. Defining rows of R as r_i, $< r_i, r_j >_{\mathcal{H}_1}$ is 0 for $i \neq j$ and 1 for $i = j$. These three characteristics of R ease the technical development of these methods. See Wahba (1990) for a full derivation.

The LASSO is a penalized regression method, where a linear model is fit subject to a constraint on the sum of the absolute values of the parameters. LASSO estimation has the desirable property of producing point estimates of *precisely* zero for most effects. This section presents the intuition behind LASSO estimation algebraically and geometrically.

More formally, assume a vector of n outcomes y, an $n \times m$ matrix of observed covariates X. The matrix X may contain main or interactive effects, as driven by theory or common practice. Assume the $m \times 1$ vector $\beta = [\beta_j]$ of parameters associated with X, the LASSO estimator is defined as the solution to the minimization problem:

$$\hat{\beta}^{LASSO} = \underset{\beta}{\operatorname{argmin}}(y - X\beta)'(y - X\beta) + \lambda_1 \sum_{j=1}^{m} |\beta_j| \tag{4.7}$$

Several similarities with the smoothing spline are apparent. First, there is a nonnegative complexity measure of the form $\Omega_{LASSO}(\beta) = \sum_{j=1}^{m} |\beta_j|$. Complexity of the LASSO model is measured by "how much β" the model has, rather than the total "curviness" of the fit. The tuning parameter λ_1 serves the same basic role as that with splines, where $\lambda_1 = 0$ returning the least squares estimate from regressing y on X, while a sufficiently large λ_1 returns $\hat{\beta}_{LASSO} = 0$.

The algebraic intuition is most apparent when the columns of X are uncorrelated. Let $\hat{\beta}^o$ be the least-squares estimates of β and $(x)_+$ denote $x \cdot I(x > 0)$. In the case where the columns of X are uncorrelated, the LASSO estimator can be written (see Tibshirani 1996: 269):

$$\hat{\beta}_j^{LASSO} = \hat{\beta}_j^o \left(1 - \frac{\lambda_1}{|\hat{\beta}_j^o|}\right)_+ \tag{4.8}$$

The LASSO estimator shrinks least squares estimates greater than λ_1 towards zero by factor $1 - \lambda_1/|\hat{\beta}_j^o|$. Covariates with least squares estimates less than λ_1 are estimated as zero. For non-orthogonal design, the LASSO solution proves intractable, because the penalty $\sum_{j=1}^{k} |\beta_j|$ is not differentiable at $\beta_j = 0$; yet the general insight that a singularity in the penalty induces sparsity in the coefficients carries through (Scholkopf and Smola, 2001).

The LASSO carries an informative geometric interpretation. LASSO estimation can be viewed as placing a constraint on a residual sum of squares (RSS), with a solution indicated where the hyperellipse $RSS = q_{LASSO}$ is tangent to the constraint. The standard form of the LASSO estimator, and a corresponding smoothed estimator,[2] is given below:

$$\hat{\beta}^{LASSO} = \underset{\beta}{\operatorname{argmin}}(y - X\beta)'(y - X\beta) \text{ subject to } \sum_{j=1}^{k} |\beta_j| \leq q_{LASSO} \tag{4.9}$$

$$\hat{\beta}^{smooth} = \underset{\beta}{\operatorname{argmin}}(y - X\beta)'(y - X\beta) \text{ subject to } \sum_{j=1}^{k} \beta_j^2 \leq q_{smooth} \tag{4.10}$$

The geometric interpretation is made clear in figure 4.2. Consider the case with only two parameters, (β_1, β_2) and associated least squares coefficients $(\hat{\beta}_1, \hat{\beta}_2)$. In this case, the ridge constraint is the circle $\hat{\beta}_1^2 + \hat{\beta}_2^2 = q_{spline}$. The LASSO constraint, in contrast, is the square $|\hat{\beta}_1| + |\hat{\beta}_2| = q_{LASSO}$. The confidence (Scheffe) ellipse is centered at $(\hat{\beta}_1, \hat{\beta}_2)$, and its shape is governed by $cov(\hat{\beta}_1, \hat{\beta}_2)$. For a given value of k_1 or k_2, the minimizer to the loss function occurs where the confidence (Scheffe) ellipse is tangent to the constraint. The ellipse will hit the smoothing constraint at a point where neither coefficient is zero. The ellipse, though, is

[2]This is the constraint used in random effects models, smoothing splines, ridge regression, or through assuming a normal prior over the coefficients. The resulting estimates differ in interpretation, based off whether β is assumed random or fixed, but the optimization is the same.

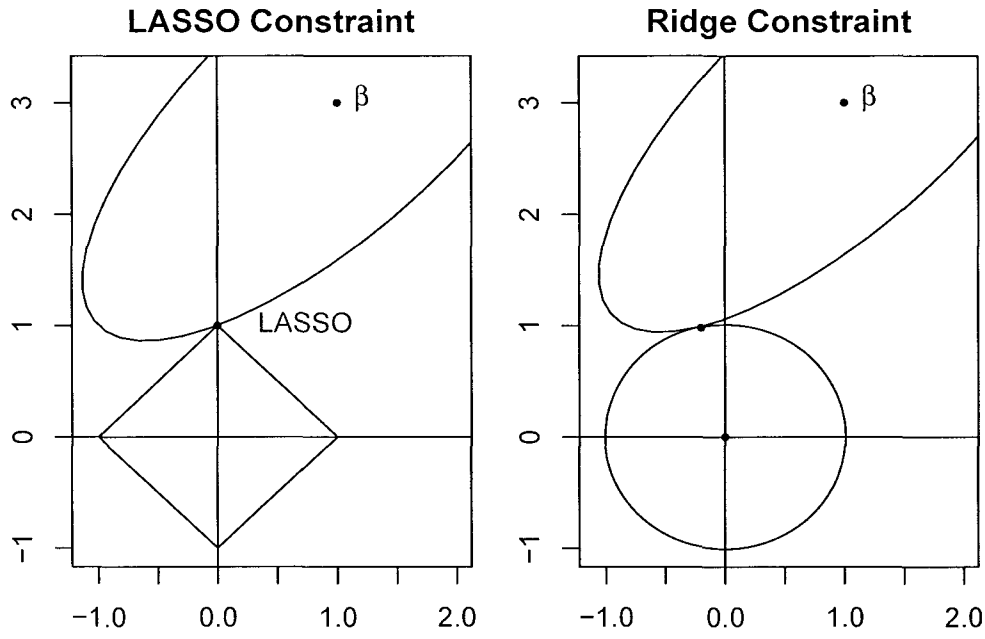

Figure 4.2: A geometric interpretation of the how the LASSO penalty produces point estimates of zero. The penalized estimates in each case are found by expanding the ellipse until it is tangent to the constraint, a diamond for the LASSO and a circle for the spline. The ellipse will hit the smoothing constraint at a point where neither coefficient is zero. The ellipse, though, is likely to hit the square at a corner, setting some of the estimates to zero.

likely to hit the square at a corner, setting some of the estimates to zero. Thus in practice, the LASSO estimator is a powerful variable selection mechanism.

4.3 The Proposed Method

The proposed mixed penalty model combines the two approaches above, through simultaneously fitting a smooth curve while selecting from a set of known covariates. The target function consists of three components: a linear trend, a smooth trend, and some subset of all possible jurisdiction-specific effects. The methods smooths over the geographic trend, while selecting the most relevant jurisdiction-specific effects. The loss function is optimized subject to both a spline constraint and a LASSO constraint.

4.3.1 The Model

Assume a vector of n observed outcomes, y. The vector y is assumed additive in three terms. The first is an unpenalized component, a matrix S with columns consisting of latitude, longitude, controls, and in intercept. The second is a matrix X, an $n \times m$ matrix of jurisdiction effects. For example, with 49 jurisdiction, and an intercept and two covariates modeled for each jurisdiction, X would contain $147 = 49 \times 3$ columns, i.e. $m = 147$. Parameters in this component are selected, so they are placed under a LASSO constraint. The final component is a dense $n \times n$ matrix R, the thin plate spline reproducing kernel, as described in section 4.2.1. Given a vector of n independent, symmetric, equivariant disturbances, ϵ, the model for y is

$$y = Sd + Rc + X\beta + \epsilon \tag{4.11}$$

To avoid overfitting, a smoothing constraint is placed over $c'Rc$, controlling the complexity of the resultant fit. In order to select elements of β, a LASSO constraint is placed over these parameters.

For a given (λ_1, λ_2), adding the constraints to the generates a mixed penalty function of the following form:

$$\{\hat{c}_{mixpen}, \hat{d}_{mixpen}, \hat{\beta}_{mixpen}\} = \tag{4.12}$$

$$\underset{c,d,b,\lambda_1,\lambda_2}{\operatorname{argmin}} (y - Sd - Rc - X\beta)'(y - Sd - Rc - X\beta) + \lambda_1 \sum_{j=1}^{m} |\beta_j| + \lambda_2 \mathbf{c}' R \mathbf{c} \tag{4.13}$$

4.3.2 The Algorithm

The algorithm proceeds in three steps. First, a value of (λ_1, λ_2) is assumed. Given these values, the model is fit. Finally, the fit is evaluated by an external criterion, a Bayesian Generalized Cross-Validation statistic. An alternating line search is done to find the optimal values of the tuning parameters that optimize this criterion.

4.3.2.1 Scaling the covariates

LASSO estimation requires scaling the covariates under the selection constraint. Following standard practice, each column of X is given a standard deviation of one. Columns of X that are jurisdiction-specific indicator (dummy) variables are left uncentered, so most values are left at zero.

Columns of X that are interactions between a jurisdiction-specific effect and a continuous covariate are scaled so that the covariate takes a value of zero for every observation outside the jurisdiction, and observations within the jurisdiction are centered. This occurs in three steps. First, the continuous covariate is de-meaned and given standard deviation one. Second, the standardized indicator variable and the continuous component are multiplied together. Finally, the elements of the covariate corresponding with the given jurisdiction are centered.

4.3.2.2 Estimating the coefficients, for a fixed value of the tuning parameters

Estimating coefficients for a given value of (λ_1, λ_2) is done in two stages. First, the matrix $A(\lambda_2)$ that projects y onto its spline fitted vaues is calculated. Second, a LASSO fit is done

on the residuals of fitting y and X with a spline, i.e. the LASSO is fit using $y^* = (I - A(\lambda_2))y$ and $X^* = (I - A(\lambda_2))X$.

R is an $n \times n$ matrix, and computationally difficult to invert. Instead, a random subset of the knots is chosen of size n_0. Gu (2002) recommends $n_0 = \text{ceiling}(\max(30, 10 \cdot n^{2/9}))$. The simulations and analyses use twice this number of knots, so $n_0 = \text{ceiling}(\max(60, 20 \cdot n^{2/9}))$. Let R_{coef} denote the $n \times n_0$ submatrix of R with columns corresponding to selected knots, and let R_{kern} denote the $n_0 \times n_0$ submatrix of R_{coef} with rows and columns corresponding to selected knots.[3] Define M as the concatenation of S and R_{coef}, $M = [S|R_{coef}]$, and V as the $(k + n_0) \times (k + n_0)$ matrix

$$V = \begin{pmatrix} 0 & 0 \\ 0 & R_{kern} \end{pmatrix} \tag{4.14}$$

Let the superscript "$-$" denote the Moore-Penrose generalized inverse. The projection matrix, $A(\lambda_2)$, is defined as

$$A(\lambda_2) = M(M'M + \lambda_2 V)^- M', \tag{4.15}$$

Next, take the residuals from applying the projection matrix to y and X, generating

$$y^* = (I - A(\lambda_2))y \tag{4.16}$$

$$X^* = (I - A(\lambda_2))X \tag{4.17}$$

Finally, solve the LASSO problem

$$\hat{\beta}^{LASSO}_{\lambda_1,\lambda_2} = \underset{\beta}{\text{argmin}}(y^* - X^*\beta)'(y^* - X^*\beta) + \lambda_1 \sum_{j=1}^{m} |\beta_j| \tag{4.18}$$

[3] R_{kern} is required to be positive semi-definite, which holds in theory but not always in practice. Following common practice, R_{kern} is passed through a filter that sets all of its negative eigenvalues to zero.

Fitted values can be calculated as $\hat{y} = A(\lambda_2)y + X^*\beta$ and residuals as $y - \hat{y}$.

4.4 The Fit Criterion

The proposed method requires the selection of multiple tuning parameters. To evaluate the fit at a given value of $\{\lambda_1, \lambda_2\}$, a Generalized Cross-Validation statistic (GCV) is fit (Wahba, 1990), in order to balance model fit against model dimensionality. The estimated degrees of freedom from the linear and spline components is taken as the trace of $A(\lambda_2)$, i.e. $edf_{spline} = tr(A(\lambda_2))$. The number of non-zero coefficients provides an unbiased estimate of the dimensionality of a LASSO model (Zou *et al.*, 2007), $edf_{LASSO} = \sum_{j=1}^{m} I(\hat{\beta}_j^{LASSO} \neq 0)$. Given a sample size of n and model dimensionality of k, the GCV statistic is

$$GCV_{\lambda_1,\lambda_2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\beta)^2}{(1 - \frac{edf_{spline}+edf_{LASSO}}{n})^2} \tag{4.19}$$

The GCV balances the residual sum of squares, in the numerator, against the model dimensionality, in the denominator. A more complex model will fit the sample better, decreasing the numerator, but will also decrease the denominator.

The proposed method finds the ideal fit by minimizing a criterion closely related to the GCV. GCV statistics are known to be inconsistent for model selection, when the model space is finite(Shao, 1997). To adjust the GCV to variable selection, I propose a Bayesian GCV (BGCV) statistic of the form

$$BGCV_{\lambda_1,\lambda_2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\beta)^2}{(1 - \frac{1}{2}\cdot\frac{2\cdot edf_{spline}+\log(n)\ edf_{LASSO}}{n})^2} \tag{4.20}$$

Intuitively, the 2 in front of edf_{spline} in equation 4.20 accounts for the asymptotic agreement between the AIC and GCV (Shao, 1997). To convert the statistic from a GCV to a

BGCV, 2 is replaced by $\log(n)$, which is asymptotically consistent for model selection. The adjustment is admittedly ad hoc, but it is designed to mimic a shift from an AIC to a BIC statistic. For a similar adjustment to a GCV statistic, see (Shi *et al.*, 2006). Simulations show that the BGCV maintains a reasonable discovery rate and a low false discovery rate.

4.4.1 Search Strategy

The search strategy consists of a series of alternating line searches across a broad range of the tuning parameters. First, λ_1 is fixed at a large value, (exp(25)). Next, λ_2 is evaluated along the set $\log(\lambda_V) \in \{-15, -14, \ldots, 10\}$, with the value producing the smallest GCV statistic selected. Given the current estimate of λ_V, λ_Z is evaluated along the set $\log(\lambda_V) \in \{-15, -14, \ldots, 10\}$. The λ_V that produces the smallest GCV statistic is selected. We alternate in a line search between the two parameters to convergence at a given precision. After convergence at a given precision, the radius is decreased, and the precision increased. The process is repeated to a precision of .0001.

4.5 Simulations

Simulations are conducted in order to assess the proposed method's efficacy. The target function is generated over a grid, and consists of a smooth curve with either zero or two square-specific effects. This function is designed to mimic a scenario in which a social outcome varies smoothly across a geography, but there may be some jurisdiction-specific effects present. Nearby observations may be correlated, due to geographic proximity, but there may also be discrete shifts at known borders.

To set up the simulations,the grid is drawn over the range $\{0, 10\} \times \{0, 10\}$, with twenty-five 2×2 squares. Samples ranging between 900 and 3,600 are drawn uniformly across this range. The systematic smooth and grid-effect components of the target function are

$$\eta_{smooth}(x, y) = \sin\left(\frac{(x+y)^2 * \pi}{500}\right) + \sin\left(\frac{(x-y)^2 * \pi}{200}\right) \tag{4.21}$$

$$\eta_{jump}(x, y) = I(2 < x < 4, 2 < y < 4) - 2 \cdot I(6 < x < 8, 4 < y < 6) \tag{4.22}$$

The target function is illustrated in figure 4.3. It is comprised of a smooth curve evaluated over the grid, with one square raised and one dropped down. The effect sizes of $+1$ and -2 are selected to be approximately 1 and -2 times the residual standard deviation. Points are generated uniformly across the square, and independent noise from a t-distribution on ten degrees of freedom is added to the systematic component. Because the possible location of these breaks are known, the X matrix has n rows and 25 columns, i.e. $m = 25$, with each column a dummy variable. The linear component, S consists of an intercept, x, and y. The goal is to fit a smooth curve over the smooth component while simultaneously selecting the grid effects. The following six simulations are executed at sample sizes of

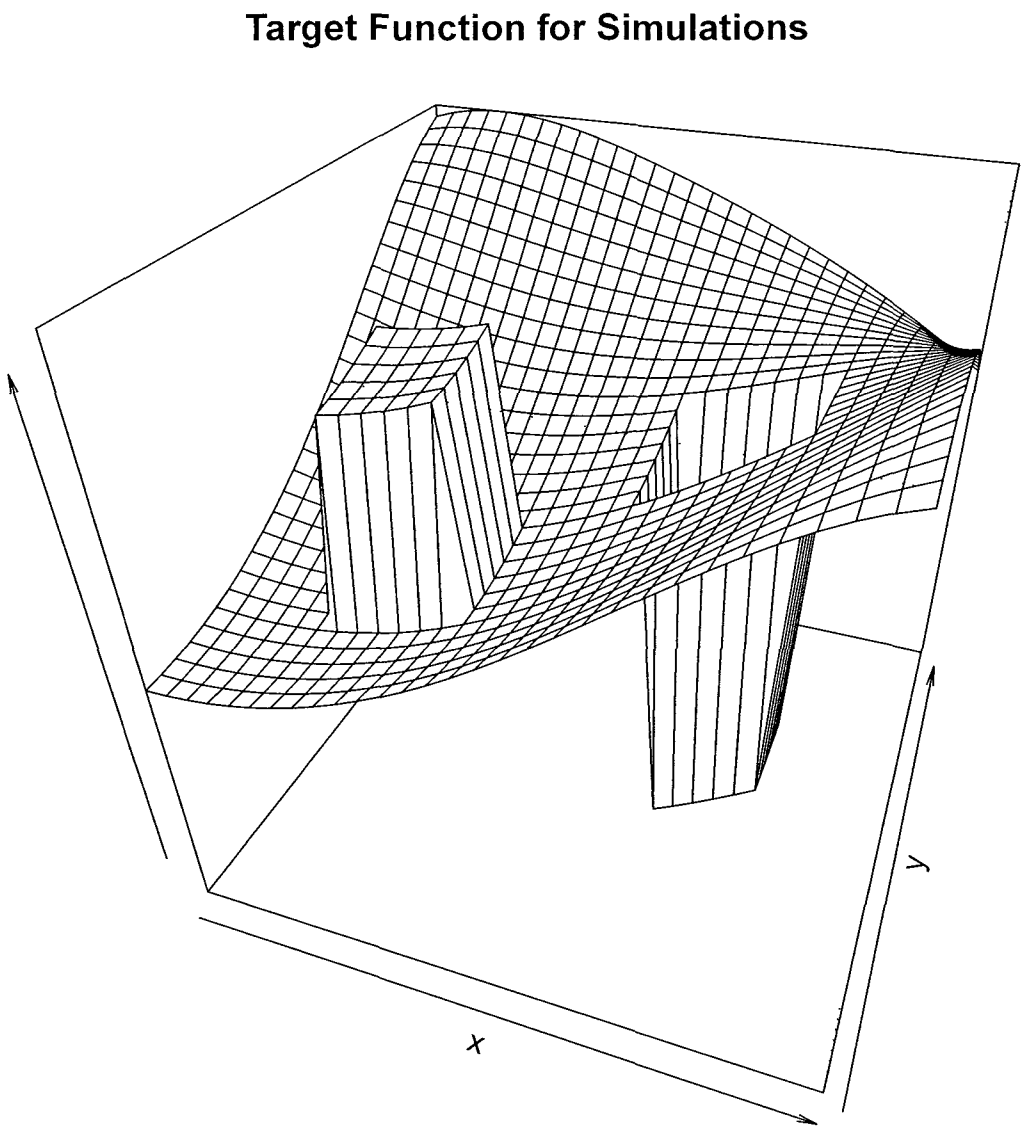

Figure 4.3: The target function for the simulations. The function is smooth, except for effects that are within two of the grid squares. This function is designed to mimic a scenario where a social outcome varies smoothly across a geography, but there may be some jurisdiction-specific effects present. Nearby states may be correlated, due to geographic proximity, but there may also be discrete shifts at known borders.

$n \in \{15^2, 30^2, 45^2, 60^2\}$:

$$\textit{Model Specifications:} \tag{4.23}$$

$$sim_{jump}(x_i, y_i) = \eta_{smooth}(x_i, y_i) + \eta_{jump}(x_i, y_i) + u_i \tag{4.24}$$

$$sim_{nojump}(x_i) = \eta_{smooth}(x_i, y_i) + u_i \tag{4.25}$$

$$\textit{Variance Specification:} \tag{4.26}$$

	Without Grid Effects, Spline model properly specified			With Grid Effects, Fixed effects model properly specified		
n	Proposed	Spline	Spline + Fix Eff	Proposed	Spline	Spline + Fix Eff
225	0.208	0.192	0.353	0.351	0.434	0.373
900	0.122	0.128	0.215	0.204	0.344	0.211
2025	0.085	0.090	0.133	0.132	0.304	0.140
3600	0.069	0.070	0.106	0.104	0.261	0.110

Table 4.1: Root mean square difference between the true curve for the proposed method and its competitors. In cases without grid effects, the method performs comparably to a smoothing spline and better than a smoothing spline with fixed effects. In cases with grid effects, the method performs comparably to a smoothing spline without fixed effects, and dominates a smoothing spline.

Simulations were executed 100 times, with the same set of data fit with the proposed method, a thin plate smoothing spline, and a thin plate smoothing spline with fixed effects for each column of X. Splines are fit using function `ssanova` in R library `gss`, using a REML estimate of the smoothing parameter.

4.5.1 Simulation Results

The simulations demonstrate the proposed method's utility. Table 4.1 shows the root mean square (RMS) difference between the estimated and true curve for each method. In the absence of grid effects, the method performs like a smoothing spline. In the presence of grid effects, it performs as well as a fixed-effects specification, even though the latter estimates many more non-zero coefficients. Without grid effects, the loss is indistinguishable from a spline; with grid effects, the loss is indistinguishable from a fixed effects specification. The proposed method appears to navigate well between these two extremes.

In the simulation with no grid effects, the proposed method selected no effects, for a false discovery rate of zero. The remainder of this analysis focuses on the simulations where there

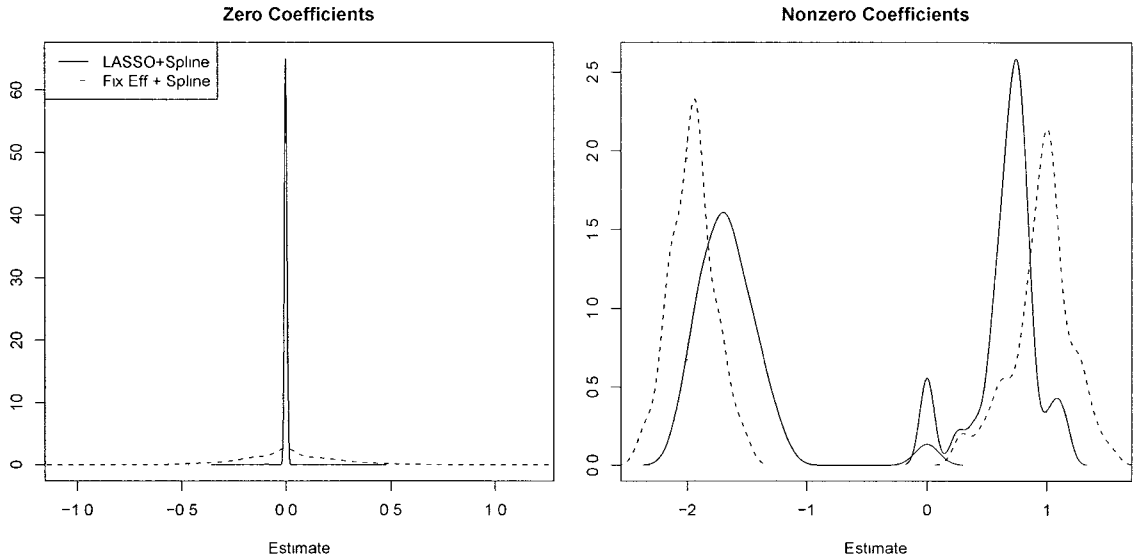

Figure 4.4: The distribution of estimated coefficients for effects that are in truth zero (left) and in truth non-zero (right) in the simulation with grid effects. The true parameter values are 1 and -2, and are indicated with vertical lines. The method sets the magnitude zero effects correctly 97.3 percent of the time, and estimates nonzero magnitude effects with the correct sign 81 percent of the time. Just as importantly, it *never* produces an estimate of the wrong sign for effects that are, in truth, nonzero.

were grid effects. Figure 4.4 illustrates the selection properties for the set of simulations where there were grid effects. The figure illustrates the distribution of the estimates of the parameters that are in truth zero (left) and nonzero (right) . When the true parameter value is zero, the proposed method produces a nonzero coefficient estimate only 3.7 percent of the time, producing a distribution that is much tighter around zero than the fixed effects estimate.

For the estimates with a true value of either 1 or -2, the fixed effects estimate are unbiased, while the LASSO estimates are shrunk towards zero. This is an illustration of the well-known "bias-variance" trade-off. In this case, the proposed method biases estimates in a particularly useful manner, setting most to zero.

The simulations demonstrate the utility of the method on reasonable sample sizes, ranging from hundreds to thousands of observations. The proposed method can select both nonzero and zero estimates with reasonable accuracy, and when there are in truth no effects, the method performs comparable to smoothing splines. The proposed method presents a sound way to fit a model with both a smooth component and a small number of jumps.

4.6 US 2008 Presidential Election Results: Red States and Blue States

As with pundits who like to "slice-and-dice our country into Red States and Blue States (Obama, 2004)," political scientists actively study state-level partisanship effects. As the United States selects its president through an Electoral College, providing a winner-take-all race for a slate of electors based on each state-level result, the question is of particular substantive interest. Popular vote is aggregated within known state lines, and these state-level results determine the assignation of all of the state's electors to a single candidate.[4] Whether something is the matter with Kansas (Frank 2004, Bartels 2006), explaining how the effect of income and population density on vote choice qualitatively differs between "red" and "blue" states (Gelman *et al.*, 2005), or typologies of state level "cultures (Elazar, 1984)," states provide a useful level of analysis. This section analyzes county-level vote returns from the 2008 presidential contest.

The dependent variable throughout this section is the log-odds ratio of county-level vote returns for Barack Obama versus John McCain. The data are from each county in the continental United States ($n = 3109$), with McCain and Obama accounting for 98.7 percent of

[4]With the exception of Maine and Nebraska.

the total popular vote. The number of counties in each state range between 1 for Washington, D.C. and 254 for Texas, with a mean of 63.4 and median of 64. The total number of votes cast in each county range between 79 for Loving, Texas and 2,818,964 for Los Angeles, California. McCain won the majority of counties (73%), while Obama won larger counties; the correlation between Obama's percent of the vote and the log population cast, weighted by total number of votes cast, was .56.

The primary stylized fact that I explore is America's urban-rural divide: urban, dense areas overwhelming supported Obama, while rural, sparse areas supported McCain. I present two models of increasing complexity. The first contains only an intercept for each continental state and Washington, D.C., so the X matrix is composed of 49 indicator variables, and $m = 49$. The second contains an intercept, a population effect, and an income effect, and so X has three times as many covariates, setting $m = 147$.

Results from the two are in table 4.2. I find two sets of results. The first set come from the intercept model. The second come from the model with effects for within-state heterogeneity in income and population. The results from the intercept model are rather unsurprising. Arizona, the home state of candidate McCain, and the known "red" states of Oklahoma, Texas, and Utah were discovered. Maryland and New York, both traditionally "blue," were uncovered. California, among the "bluest" of states, was not uncovered–and the saturated model explains why.

The second model, which includes state-specific effects for log population and log median income by state, casts unique insight into the election results. First, California and New

	Intercept only	Intercept, population, and income effects		
State	Intercept	Intercept	Population	Median income
Alabama	0	-0.057	0	0
Arizona	-0.139	0	-0.046	0
California	0	0.121	-0.123	0.084
Colorado	0	0.050	0	0
Connecticut	0	-0.011	0	-0.013
Delaware	0	0.007	0	0
District of Columbia	0	0.075	0	0
Illinois	0	0.041	0	0
Iowa	0	0.056	0	0
Kentucky	0	-0.042	0	0
Maryland	0.161	0.021	0.098	0
Nevada	0	0.024	0	0
New Hampshire	0	0	0	-0.011
New Jersey	0	-0.009	0	-0.026
New Mexico	0	0.099	0	0
New York	0.087	0	0.268	-0.212
Ohio	0	-0.039	0	0
Oklahoma	-0.066	-0.098	0	0
Oregon	0	0.082	0	0
Pennsylvania	0	-0.038	0.034	0
South Carolina	0	0	0	-0.032
Tennessee	0	-0.057	0	0
Texas	-0.051	-0.017	0	-0.039
Utah	-0.106	-0.081	0	0
Vermont	0	0.036	0	0
Virginia	0	0.006	0	0
Washington	0	0.038	0	0
Wisconsin	0	0.025	0	0

Table 4.2: Results for state level analyses. Positive coefficients indicate a pro-Obama effect; negative coefficients indicate a pro-McCain effect. Results are on a log-odds scale. The first column identifies well-known "red" states Texas, Utah, and the home state of John McCain, Arizona. The known blue states of New York and Maryland are identified. The model with state-specific income and population effects gives a subtler picture. California had a blue effect, but Barack Obama performed worse in high-population areas, and better in richer areas. In New York, Barack Obama performed better in high-population areas and worse in high-income areas. In Pennsylvania, Obama performed well in populated areas, Philadelphia and Pittsburgh, but poorly elsewhere in the state.

York both went heavily for Barack Obaama, but they did so in different ways. California has a strong intercept effect in the second model, but not the first, because it was masked by within-state heterogeneity. Barack Obama did better in wealthier areas, but worse in more populated areas, than the national trend. In New York, the opposite result holds. Barack Obama performed will in more-populated areas, but worse in wealthier areas in California. California and New York are both "blue" states, but they are blue in different ways.

In Texas, John McCain performed well on average, and as the second model shows, he performed better in wealthier areas. In Pennsylvania, John McCain performed well on average, but Barack Obama captured the high-density areas of Philadelphia and Pittsburgh at a level above the national trend. Pennsylvania was a hotly-contested state, with Barack Obama focusing on boosting turnout in urban areas, and John McCain focusing through the rest of the state. This strategy is reflected in the data.

Finally, the results differ from those of Gelman *et al.* (2005). One of their central findings separates cross-state from within-state variation. They show that states with a higher average income are more likely to support the Democratic presidential candidate, but wealthy citizens within these states are more likely to support the Republican candidate. The reason for the divergence between the two sets of findings is two-fold. First, Gelman, et al., analyze individual-level data, while I use county-level data. Second, much of what Gelman, et al., discover may be regional effects, rather than an effect particular to a given state. They do not account for geographic correlation, which may be driving their results.

4.7 The Geographic Distribution of GDP across Africa

The G-Econ project, undertaken by William Nordhaus and colleagues, has produced a dataset that estimates GDP at each degree of latitude and longitude across the globe, as well as a battery of geographic and climatological covariates.[5] Recent debates have focused on the role of long-run historical institutional trends in explaining cross-national differences in GDP (Acemoglu, et al. 2001; McArthur and Sachs 2001). These explanations, though, predicate some relation that differs from one side of a national border to the other. The statistical analyses in Acemoglou, et al. and McArthur and Sachs include country indicators; this misses the natural correlations due to geographic proximity between- and within-countries. As opposed to traditional fixed-effct specifications, the proposed method accounts for within-country variation and a geographic trend, while identifying an effect that ends at political boundaries.

Considering the within- and between-country effects are crucial. Local geographic factors can be used as exogenous instruments (Miguel *et al.*, 2004), and are crucial in explaining both economic growth and its dampening effect on conflict. Instrumental variable analysis, though, requires a proper specification of the causal mechanism and estimates an average causal effect. This precludes causal heterogeneity, by construction, and rests heavily on the presumption that the causal mechanism has been properly modeled. The method used here is pre-causal, in that only correlations are uncovered, but it can point researchers towards specific areas that merit further attention. As Todd Moss states, "Indeed, the most glaring

[5]Details and complete documentation available at http://www.gecon.yale.edu. Accessed March 9, 2009.

trend is the divergence among African countries facing opposing economic and political trajectories (2007, 6)," and a statistical analysis should take this divergence into account.

In this section, I fit a model with state-level effects, trying to identify sustained mean-level effects corresponding with either state boundaries or former colonizing power. The G-Econ data provides information on GCP (gross cell product, estimated production at that location, n=3306), average rainfall, elevation, and temperature. All data were collected between 1985 and 1990. The dependent variable for each model is the log of GCP per capita. The square root of rainfall, elevation, and temperature variables were taken to reduce skew. Observations are weighted by 1990 population. Two specifications are fit. The first includes an intercept for each country. The second includes an intercept and linear term in population for each effect.

I focus on interpreting the states with largest effects in the intercept. I find three different type of effects, after accounting for local conditions and geographic correlatino. First, oil produces have a higher state-effect. Second, states with some semblance of political pluralism also have a positive effect. Third, a longstanding territorial conflict or guerrilla movement decreases state-level GDP.

The easiest effects to explain are the strong positive effects for Nigeria and Sudan. Both are rich oil-producers, and the proposed method identified these countries as such. These effects serve primarily as a validity check rather than any unique insight. The second model shows that, in Nigeria, the effect is focused primarily in less-dense areas. Whether the

Country	Effect size
Algeria	-0.024
Cameroon	0.139
Guinea Bissau	-0.011
Kenya	0.251
Lesotho	-0.028
Libya	0.059
Mali	-0.049
Mauritania	-0.059
Morocco	-0.244
Mozambique	-0.292
Namibia	-0.007
Niger	-0.029
Nigeria	0.354
Senegal	-0.059
South Africa	0.147
Sudan	0.147

Table 4.3: Results for the African analysis. Positive coefficients indicate higher GDP per capita. Results are on a log-odds scale. Positive effects were identified for oil produces (Sudan and Nigerian) and states with some semblance of political pluralism (Cameroon and South Africa). States with a negative identified effect were engaged in either a longstanding separatist movement (Morocco) or guerrillas actively disrupting the economy (Mozambique).

additional production is a "resource curse" is hotly debated in the literature, but at the least, oil does appear to increase state-level production (Ross, 1999).

In 1990, Kenya, and South Africa stood out in their relatively peaceful politics, especially in relation to Burundi and Mozambique. In 1990, South Africa was the only African country to have had a peaceful, election-driven transition (Nugent 2004, 369). Similarly, the 1980's found Kenya under President Moi with, if not a democratic regime, at least one demonstrating a resurgence in political pluralism Ngunyi and Gathiaka (1993). Politics in Cameroon in the 1980's involved several attempts to unseat still-President Paul Biya. Though the

attempts were unsuccessful, and the 1992 elections showed signs of fraud, the process was peaceful. Similar to Kenya, Cameroon was a government with a strong political leader, but there was at least some room for relatively weak parts of the ruling party to peacefully, though unsuccessfully, compete in the political arena.

Morocco and Mozambique stand in stark contrast to Cameroon, Kenya, and South Africa. In 1990, Morocco was facing an open conflict with the separatist Polisario. By 1994, this dispute was "the only major unresolved colonial question," at the expense of "the expenditure of thousands of human lives (and) billions of dollars (Pazzanita, 1994)." Having begun after Morocco's invasion of the Spanish Sahara in 1975 after the International Court of Justice ruled that Morocco did not have a claim to the territory, open conflict continued until a 1991 cease fire was declared. Similarly, Mozambique in the 1980's saw the insurgent National Resistance of Mozambique (RENAMO) using guerilla methods to upset the ruling Liberation Front of Mozambique (FRELIMO). To quote Paul Nugent:

> "RENAMO rapidly evolved into a debilitating scourge in the early 1980's...The underlying aim was...to further compound the economic crisis and thereby to undermine the credibility of the government (2004, 284)."

My mixed penalty method has managed to select two different sets of countries: those with some evidence of political competition,and those where political and ethnic lines both coincide and have led to violence. After accounting for local geographic conditions, the selected countries suggest a strong relation between a peaceful democratic process and higher GCP per capita. The countries with a positive country-level effect had, to varying degrees,

some semblance of pluralist political competition, while the countries with negative country-level effects faced internal guerilla insurgence and mass slaughter in lieu of peaceful political negotiation.

4.8 Conclusion

Political scientists are increasingly using spatial data in order to identify effects that correspond with jurisdictional boundaries. Identifying effects while also accounting for geographic correlation is the central problem in these analyses. A second problem is that the number of estimated coefficients grows rapidly. Researchers may want to fit models with dozens to hundreds of coefficients, including an intercept and several linear trends for each state. The proposed method accomplishes both goals simultaneously: modeling local geographic correlation while selecting only the most relevant covariates.

The proposed method integrates smoothing splines with LASSO variable selection, a novel combination in the political science and statistical literatures. The chapter also serves as a gentle introduction to these two methods, communicating the intuition behind these methods and their implementation.

Simulations and two different analyses illustrate the proposed method's utility and insights. The simulations reveal that the method is powerful, correctly identifying effects that are present, and rarely misleading, with a low false positive rate. The analyses illustrate insights that can be gleaned from the method. The positive relationship between median county income in two states, New York and Massachusetts, and negative relationship in two states, Washington and California, can help open up new avenues of research in the study of

political behavior. The analysis of GDP in Africa revealed three different sets of states. The first set is noticeable for its high level of civil society, and the method uncovers higher levels of economic production in those states. The uncovered countries with lower levels of civil society are associated with less growth. The third set of countries were facing an economic shock due to the drop in oil in one case and a drought in the other. This uncovered relation between civil society and growth is only correlative. Correlation is not causation, yet the two are highly correlated. The proposed method uncovers subtle, unexpected correlations that would have otherwise lain undiscovered.

Finally, the proposed method illustrates how data-driven variable selection methods can be used through the field. Often, we have little, weak, or contradictory a priori theory to guide us in selecting variables and crafting hypotheses about the expected direction of an effect. In these scenarios, where a researcher simply wants to discern what hypotheses are most suggested by the data, the proposed method and other related data-driven variable selection estimators afford the researcher precisely that opportunity.

Bibliography

Acemoglu, D., Johnson, S., and Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *The American Economic Review* **91**, 5, 1369–1401.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association* **91**, 434, 444–455.

Assmann, S. F., Pocock, S. J., Enos, L. E., and Kasten, L. E. (2000). Subgroup analysis and other (mis)uses of baseline data in clinical trials. *The Lancet* **355**, 9209, 1064–1069.

Bartels, L. (2006). What's the matter with What's the matter with Kansas? *Quarterly Journal of Political Science* **1**, 201–226.

Beck, N. and Jackman, S. (1997). Beyond linearity by default: Generalized additive models. *American Journal of Political Science* **42**, 2, 596–627.

Beck, N., Jackman, S., and Rosenthal, H. (2006). Presidential approval: the case of George W. Bush. Working Paper.

Beck, N., King, G., and Zeng, L. (2000). Improving quantitative studies of international conflict: A conjecture. *American Political Science Review* **94**, 1, 21–36.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**, 1, 289–300.

Berry, W. D. and Baybeck, B. (2005). Using Geographic Information Systems to Study Interstate Competition. *American Political Science Review* **99**, 04, 505–519.

Bickel, P., Li, B., Tsybakov, A., Geer, S., Yu, B., Valds, T., Rivero, C., Fan, J., and Vaart, A. (2006). Regularization in statistics. *TEST. An Official Journal of the Spanish Society of Statistics and Operations Research* **15**, 2, 271–344.

Bivand, R. S., Pebesma, E. J., and Gómez-Rubio, V. (2008). *Applied Spatial Data Analysis with R*. Springer, New York. ISBN 978-0-387-78170-.

Bradley, P. and Mangasarian, O. L. (1998). Feature selection via concave minimization and support vector machines. In *Machine Learning Proceedings of the Fifteenth International Conference*, 82–90. Morgan Kaufmann.

Brady, H. E. and Collier, D. (2004). *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, MD: Rowman & Littlefield Publishers,.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics* **24**, 6, 2350–2383.

Breiman, L. (2001). Random forests. *Machine Learning* **45**, 5–32.

Calderia, G. A. and Zorn, C. J. W. (1998). Of time and consensual norms in the Supreme Court. *American Journal of Political Science* **42**, 3, 874–902.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *Annals of Applied Statistics* **4**, 1, 266–298.

Cleveland, W. S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association* **74**, 829–836.

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist* **49**, 12, 997–1003.

Cole, S. R. and Stuart, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *American Journal of Epidemiology* **172**, 1, 107–115.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2008). Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics* **90**, 3, 389–405.

Davison, A. C. (1992). Treatment effect heterogeneity in paired data. *Biometrika* **79**, 3, 463–474.

Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association* **94**, 1053–1062.

Efron, B. (1986). Why isn't everyone a Bayesian? With discussion and a reply by the author. *The American Statistician* **40**, 1, 1–11.

Efron, B., Burman, P., Denby, L., Landwehr, J. M., Mallows, C. L., Shen, X., Huang, H. C., Ye, J., Ye, J., and Zhang, C. (2004a). The Estimation of Prediction Error: Covariance Penalties and Cross-Validation [with Comments, Rejoinder]. *Journal of the American Statistical Association* **99**, 467.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004b). Least angle regression. *The Annals of Statistics* **32**, 2, 407–499.

Elazar, D. (1984). *American Federalism: A View from the States* New York: Harper and Row.

Erikson, R. S , MacKuen, M., and Stimson, J. A. (2002). *The Macro Polity.* Cambridge Studies in Public Opinion and Political Psychology. Cambridge University Press.

Fan, J and Lv, J. (2009). A selective overview of variable selection in high dimensional feature space (invited review article). *Review of Scientific Instruments* **81**, 12, 44.

Ferrari, D., Pistoresi, B., and Salsano, F. (2009). Political institutions and central bank independence revisited. Technical Report 0616.

Frangakis, C. (2009). The calibration of treatment effects from clinical trials to target populations. *Clinical Trials* **6**, 2, 136–140.

Frank, T. (2004). *What's the Matter with Kansas? How Conservatives Won the Heart of America.* New York: Metropolitan Books.

Freund, Y. and Schapire, R. E. (1999). A short introduction to boosting. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, 1401–1406. Morgan Kaufmann.

Friedman, M. (1953). *Essays in Positive Economics.* University of Chicago Press, Chicago, IL.

Gail, M. and Simon, R. (1985). Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* **41**, 2, 361–372.

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics* **2**, 4, 1360–1383.

Gelman, A., Shor, B., Bafumi, J., and Park, D. (2005). Rich state, poor state, red state, blue state: What's the matter with Connecticut? **1**, 2006.

Gelman, A. and Weakliem, D. (2007). Of beauty, sex, and power: statistical challenges in estimating small effects. Tech. rep., Columbia University.

Gerber, A. S. and Green, D. P. (2000). The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment. *American Political Science Review* **94**, 653–663.

Green, D. P. and Kern, H. L. (2010a). Detecting heterogenous treatment effects in large-scale experiments using Bayesian additive regression trees. *The Annual Summer Meeting of the Society of Political Methodology, University of Iowa* .

Green, D. P. and Kern, H. L. (2010b). Generalizing experimental results. *The Annual Meeting of the American Political Science Association, Washington D.C.* .

Green, P. (1995). Reversible jump mcmc computation and bayesian model determination. *Biometrika* **82**, 711–732.

Gu, C. (2002). *Smoothing Spline ANOVA Models*. Springer Series in Statistics. Springer.

Gunter, L., Zhu, J., and Murphy, S. A. (2011). Variable selection for qualitative interactions. *Statistical Methodology* **8**, 42–55.

Hartman, E., Grieve, R., and Sekhon, J. S. (2010). From SATE to PATT: The essential role of placebo test combining experimental and observational studies. *The Annual Meeting of the American Political Science Association, Washington D.C.* .

Hastie, T., Tibshirani, R., and Friedman, J. (2001a). *The Elements of Statistical Learning*. Springer-Verlag.

Hastie, T., Tibshirani, R., and Friedman, J. (2001b). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York.

Hesterberg, T., Choi, N. H., Meier, L., and Fraley, C. (2008). Least angle and l1 penalized regression: A review. *Statistics Surveys* **2**, 61–93.

Hill, J. L. and McCulloch, R. E. (2007). Bayesian nonparametric modeling for causal inference.

Holland, P. W. (1986). Statistics and causal inference (with discussion). *Journal of the American Statistical Association* **81**, 945–960.

Imai, K. (2005). Do get-out-the-vote calls reduce turnout?: The importance of statistical methods for field experiments. *American Political Science Review* **99**, 2, 283–300.

Imai, K. and Strauss, A. (2011). Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. *Political Analysis* **19**, 1, 1–19.

Keele, L. (2006). How to be smooth· Automated smoothing in political science. Unpublished Manuscript.

Keele, L. (2008). *Semiparametric Regression for the Social Sciences*. Wiley.

Keele, L. and Titiunik, R. (2011). Geographic boundaries as regression discontinuities. Presented at the Midwest Political Science Association, May 1, 2011.

Key, V. (1955). A theory of critical elections. *The Journal of Politics* **17**, 3–18.

Key, V. (1959). Secular realignment and the party system. *The Journal of Politics* **21**, 2, 198–210.

Kimeldorf, G. S. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications* **33**, 1, 82–95.

King, G. (1998). *Unifying Political Methodology: The Likelihood Theory of Statistical Inference.* University of Michigan Press, Ann Arbor.

Knight, K. and Fu, W. (2000). Asymptotics for Lasso-Type estimators. *The Annals of Statistics* **28**, 5, 1356–1378.

Krivobokova, T. and Kauermann, G. (2007). A note on penalized spline smoothing with correlated errors. *Journal of the American Statistical Association* **102**, 480, 1328–1337.

Lagakos, S. W. (2006). The challenge of subgroup analyses: Reporting without distorting. *New England Journal of Medicine* **354**, 1667–1669.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review* **76**, 4, 604–620.

LeBlanc, M. and Kooperberg, C. (2010). Boosting predictions of treatment success. *Proceedings of the National Academy of Sciences* **107**, 31, 13559–60.

Leeb, H. and Potscher, B. M. (2008). Sparse estimators and the oracle property, or the return of Hodges' estimator. *Journal of Econometrics* **142**, 1, 201–211.

MacKay, D. (2003). *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press.

Maki, U., ed. (2009). *The Methodology of Positive Economics.* Cambridge: Cambridge University Press.

Manski, C. F. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica* **72**, 4, 1221–1246.

Mayhew, D. (2002). *Electoral realignments: A critique of an American genre.* New Haven: Yale University Press.

McArthur, J. W. and Sachs, J. D. (2001). Institutions and geography: Comment on Acemoglu, Johnson and Robinson (2000). *National Bureau of Economic Reseach Working Paper* **No. 8114**.

Miguel, E., Satyanath, S., and Sergenti, E. (2004). Economic shocks and civil conflict: An instrumental variables approach. *Journal of Political Economy* **112**, 4, 725–753.

Moodie, E. E. M., Platt, R. W., and Kramer, M. S. (2009). Estimating response-maximized decision rules with applications to breastfeeding. *Journal of the American Statistical Association* **104**, 485, 155–165.

Moss, T. (2007). *African Development: Making Sense of the Issues and Actors.* Lynne Rienner Publishers.

Ngunyi, M. and Gathiaka, K. (1993). State-civil instituions relations in kenya in the 1980s. In P. Gibbon, ed., *Social Change and Economic Reform in Africa.* Uppsala: Nordiska Afrikainstitutet.

Nugent, P. (2004). *Africa Since Independence: A Comparative History.* Palgrave Macmillan.

Obama, B. (2004). 2004 Democratic National Convention keynote address. Delivered to the Democratic National Convention.

Park, Trevor, Casella, and George (2008). The bayesian lasso. *Journal of the American Statistical Association* **103**, 482, 681–686.

Pazzanita, A. (1994). Morocco versus Polisario: a political interpretation. *The Journal of Modern African Studies* **2**, 263–378.

Pearce, N. and Wand, M. (2006). Penalized splines and reproducing kernel methods. *The American Statistician* **60**, 3, 233–240.

Pierson, P. (2004). *Politics in Time: History, Institutions, and Social Analysis.* Princeton: Princeton University Press.

Pineau, J., Bellemare, M. G., Rush, A. J., Ghizaru, A., and Murphy, S. A. (2007). Constructing evidence-based treatment strategies using methods from computer science. *Drug and Alcohol Dependence* **88S**, S52–S60.

Poole, K. T. and Rosenthal, H. (1997). *Congress : a political-economic history of roll call voting.* Oxford University Press, New York.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Raftery, A. (1995). Bayesian model selection in social research. *Sociological Methodology* **25**, 111–63.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks.* Cambridge University Press.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 1, 41–55.

Ross, M. (1999). The Political Economy of the Resource Curse .

Rothwell, P. M. (2005). Subgroup analysis in randomized controlled trials: importance, indications, and interpretation. *The Lancet* **365**, 9454, 176–186.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics* **29**, 159–183.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology* **66**, 688–701.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics* **6**, 34–58.

Rubin, D. B. (1990). Comments on "On the application of probability theory to agricultural experiments. Essay on principles. Section 9" by J. Splawa-Neyman translated from the Polish and edited by D. M. Dabrowska and T. P. Speed. *Statistical Science* **5**, 472–480.

Scholkopf, B. and Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, Cambridge, MA, USA.

Sen, A. and Srivastava, M. (1975). On tests for detecting the change in the mean. *The Annals of Statistics* **3**, 1, 98–108.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica* **7**, 221–264.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis.* Cambridge University Press.

Shi, W., Wahba, G., Wright, S., Lee, K., Klein, R., Klein, B., Shi, W., Wahba, G., Wright, S., Lee, K., Klein, R., and Klein, B. (2006). Lasso-patternsearch algorithm with application to ophthalmology and genomic data. *Statistics and Its Interface* **1**, 137–153.

Spirling, A. (2007a). Bayesian Approaches for Limited Dependent Variable Change Point Problems. *Political Analysis* 1–19.

Spirling, A. (2007b). Turning points in the Iraq conflict: Reversible jump Markov Chain Monte Carlo in political science. *The American Statistician* **61**, 4, 1–6.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., and Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* **174**, 2, 369–386.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B* **58**, 1, 267–288.

Tierney, L. and Kadane, J. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* **81**, 82–86.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer, New York.

Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory*. Springer, New York, 2nd edn.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia, PA, USA.

Wahba, G. (2002). Soft and hard classification by reproducing kernel hilbert space methods. *Proceedings of the National Academy of Sciences* **99**, 26, 16524–16530.

Ward, M. D. and O'Loughlin, J. (2002). Spatial processes and political methodology: introduction to the special issue. *Political Analysis* **10**, 3, 211–216.

Western, B. and Kleykamp, M. (2004). A bayesian change point model for historical time series analysis. *Political Analysis* **12**, 354–374.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B* **68**, 1, 49–67.

Zhang, H. H. (2006). Variable selection for support vector machines via smoothing spline ANOVA. *Statistica Sinica* **16**, 659–674.

Ziliak, S. T. and McCloskey, D. N. (2007). *The Cult of Statistical Significance : How the standard error costs us jobs, justice, and lives*. University of Michigan Press, Ann Arbor.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B* **67**, 2, 301–20.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *Annals of Statistics* **35**, 5, 2173–2192.

APPENDIX
Derivation of the BIC and Our Modified BIC

This appendix provides a derivation of the modified BIC statistic used as our stopping rule. We follow the presentation by Adrian Raftery (Raftery, 1995), while a more technical discussion can be found elsewhere (Tierney and Kadane, 1986).

Given observed data D, candidate models M_i, $i \in \{1 \dots m\}$, and associated parameters θ_i, the posterior can be written as

$$p(D|M_i) = \int p(D|\theta_i, M_i)p(\theta_i|M_i)d\theta_i \tag{A.1}$$

Letting $g(\theta_i) = \log\{p(D|\theta)p(\theta)\}$, we now expand $g(\theta)$ around the posterior mode, $\tilde{\theta}_i$:

$$g(\theta) = g(\tilde{\theta}) + (\theta - \tilde{\theta})^T g'(\tilde{\theta}) + \frac{1}{2} \cdot (\theta - \tilde{\theta})^T g''(\tilde{\theta})(\theta - \tilde{\theta}) + o(||\theta - \tilde{\theta}||^2) \tag{A.2}$$

g' and g'' denote the gradient vector and Hessian matrix of g with respect to θ. Since $\tilde{\theta}$ is the posterior mode, we know that the linear term above is zero, since $g'(\tilde{\theta}) = 0$. This leaves the approximation:

$$g(\theta) \approx g(\tilde{\theta}) + \frac{1}{2} \cdot (\theta - \tilde{\theta})^T g''(\tilde{\theta})(\theta - \tilde{\theta}) \tag{A.3}$$

Exponentiating both sides, and noting that the integral is proportional to the integral of a multivariate normal, gives the result (MacKay, 2003):

$$p(D) \approx \exp\left(g(\tilde{\theta})\right) \int \exp\left(\frac{1}{2} \cdot (\theta - \tilde{\theta})^T g''(\tilde{\theta})(\theta - \tilde{\theta})\right) d\theta \tag{A.4}$$

$$= \exp\left(g(\tilde{\theta})\right) (2\pi)^{\frac{d}{2}} |A|^{-\frac{1}{2}} \tag{A.5}$$

Assume that A approaches infinity as some nA_0, with A_0 the asymptotic information matrix. Then $\log|A|$ approaches infinity as $\log|nA_0| = \log(n^d|A_0|)$. Substituting this into the equation above, and dropping all terms that are not changing in n or i leaves:

$$\lim_{n\to\infty} \log P(D) \approx \lim_{n\to\infty} \log P(D|\hat{\theta}_i) - \frac{1}{2}\log|A| \tag{A.6}$$

$$= \lim_{n\to\infty} \log P(D|\hat{\theta}_i) - \frac{1}{2}\log(n^d|A_0|) \tag{A.7}$$

$$= \log P(D|\hat{\theta}_i) - \frac{d}{2}\log(n) \tag{A.8}$$

The first term on the righthand side above is the log-likelihood, while the second is a measure of the dimensionality of the model. BIC is an asymptotic result, balancing the tradeoff between model likelihood and dimensionality, with the model corresponding to the highest BIC generally selected. Often, it is written as *-2· log-likelihood +d · log (n)*, in which case the model with the smallest BIC would be selected. The BIC provides an estimate of the posterior probability given a uniform distribution across all candidate models. If the correct model is a candidate, it will be chosen with probability one as the sample size approaches infinity.

We modify the BIC to account for the fact that the generalized smoothing spline fits a spline to a random subset of knots, due to the difficulties involved in inverting large matrices. Assume, with sample size n, that the spline is fit to n^* knots, chosen at random. This leads to covariance matrices $|A|$ and $|A^*|$, each which approach the same $|A_0|$, as n and n^* increase.

Taking the approximations:

$$nA \approx n^* A^* \tag{A.9}$$

$$\frac{n}{n^*} A \approx A^* \tag{A.10}$$

$$\log |A^*| \approx \log \left\{ \left(\frac{n}{n^*} \right)^d |A| \right\} \tag{A.11}$$

Given that A grows as nA_0, this leaves

$$\log |A^*| \approx \log \left\{ \left(\frac{n}{n^*} \right)^d |nA_0| \right\} \tag{A.12}$$

$$= \log \left\{ \left(\frac{n^2}{n^*} \right)^d |A_0| \right\} \tag{A.13}$$

$$= 2d \log(n) - d \log(n^*) + \log |A_0| \tag{A.14}$$

Substituting equation A.14 into equation A.6 leaves our modified BIC statistic, where the first term in the sum is the divergence of the sampled data, y^*, which is sampled from the complete set of observed outcomes, y:

$$BIC(n, n^*, i) = \log P(y^* | \hat{\theta}_i) - d \log(n) + \frac{d}{2} \log(n^*) + \frac{d}{2} \log(2\pi)$$

Rather than a likelihood, we use the penalty on the non-linear part as the measure of divergence. The logic carries through identically upon noting the correspondence between the cubic smoothing splines used here and penalized regression (Pearce and Wand, 2006). Note how, as n^* gets closer to n, our modified BIC approaches the actual BIC. Simulations with $n = 100$ and $n^* = 30$ indicated that the term with $\log(2\pi)$ was necessary to maintain a small false discovery rate in such a small sample. This asymptotically negligible term did not affect the power in the larger-n simulations, though.

We use this statistic as a stopping rule throughout our analysis and simulations.