Identifying the Effects of Political Boundaries: Simultaneous Variable Selection and Curve Fitting through Mixed-Penalty Regularization

Marc Ratkovic

July 13, 2009

Abstract

Theory in political science often presupposes some effect or outcome that operates within a jurisdictional boundary. This paper presents a statistical method that identifies sustained within-state effects, while accommodating the smooth progression of geographic data. Rather than producing an estimate for each state-level coefficient, only a few are selected, and the rest estimated at zero; the gains result from fitting complex models while producing parsimonious results. The method is situated within a broad class of regularization methods, and is illustrated on three observed data sets: county level vote returns in the 2008 US Presidential election, 2006 US county level crime rates, and 1990 African GDP, estimated at each degree of latitude and longitude.
1 Introduction

Political and social outcomes can vary from one side of a jurisdictional boundary to the other. These discontinuous shifts may arise for a variety of reasons: citizen self-selection on public good provision (Tiebout 1956), path-dependent institutional and cultural factors (Pierson 2004), or simple changes in local law. Whether profound or uneventful, the magnitude of these effects are of distinct interest to political scientists. Theories of states racing-to-the-bottom, acting as policy magnets, converging or diverging on policies, or reflecting deep-seated historical trends all imply that some factor behaves differently on two separate sides of a political border. The problem is exclusively empirical. We all agree that geography matters, but the difficulty comes in systematically disentangling effects specific to the jurisdiction from effects that cross political boundaries.

This suggests a model of socioeconomic data additive in two components: a discrete effect that breaks at jurisdictional borders and a smooth component that blends across borders. In this paper, I develop a statistical method that can handle this data structure. It selects jurisdiction-specific intercepts, while accommodating the smooth progression of local data across these same borders. The curve is best illustrated graphically; figure 1 presents the type of function that my mixed penalty method is designed to handle. While existing semiparametric methods can fit a curve of this type, through fixed or random effects, the resulting estimates are often unwieldy. A model fit to the continental United States with a state-level intercept and two state-level variables would produce 144 (= 48×3) coefficient estimates beyond those used for the smooth curve. My method offers an improvement, simultaneously fitting a smooth curve and selecting important state level covariates. Rather than producing a point estimate for each variable, only a few are selected—about 10% to 45% in the datasets used here—and the rest estimated at zero; the analytical gains result from fitting complex models while simultaneously producing parsimonious results. Three observed data sets, two in the United States and one in Africa illustrate the method’s utility. Analysis of the selected coefficients lead to different insights in the underlying political
The method works through combining two existing classes of estimators. The smooth component is that of the popular method of smoothing splines (Cristiani and Shawe-Taylor 2000, see Keele 2008 for a recent introduction aimed at social scientists). “Variable selection” refers, broadly, to any method that produces point-estimates of zero for some subset of possible variables. The variable selection method used here, the Least Absolute Shrinkage and Selection Operator (LASSO), shrinks most coefficients to zero, while producing shrunken, non-zero point estimates for the remainder (Tibshirani 1996).

Variable selection with the LASSO has received much attention in recent theoretical and applied work—see Hesterberg (2008) for an overview of the field. In normal practice, variable selection and estimation occur sequentially: coefficients are estimated, and then t- or z-values are used to determine whether the coefficient has sufficient magnitude and precision to differentiate it from zero. By producing coefficient estimates of precisely zero for a majority of covariates, the LASSO folds selection and estimation into a single step. The primary methodological innovation involves simultaneously selecting one set of (jurisdiction-specific) variables while accounting for the geographic, within-jurisdiction correlation. The goal of the paper is both to develop this method as well as situate it within a broad class of regularization techniques (to be described below)——unifying it with methods both commonly used in the discipline, as well as others that might be of interest.

The statistical problem, in fact, is both endemic and unique to political science. We often presuppose a data structure of the form described here, with some effect ending at a political border. As a specific example, consider whether the United States has “red states” and “blue states.” In this case, a latent cultural variable leads to a statewide Democratic or Republican bias, which has been found to effect the within-state relationship between income and vote choice (Gelman et al. 2006). Yet, a question remains: are there red states and blue states, or simply red regions and blue regions that happen to be delineated by state lines?
The familiar map of 2008 electoral outcomes shows regions both red and blue, with states won by Barack Obama clustered on the coasts, Midwest, and Sun Belt, while John McCain captured states through the Great Plains and South (Figure 2, top). Looking at county-level data reveals a more subtle phenomenon (Figure 2, 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 jumps at state lines. The two maps in figure 2 beg the question: do regional partisan differences respect state lines?

I find two different forms of effects. First, states may have counties with, on average, some “red” or “blue” effect. Second, the effect of county population on outcome may vary within that state in a significantly different manner. The average (intercept) effect produces nonsurprising results: New York, California, Maryland, and the District of Columbia are “blue,” while Oklahoma, Texas and Utah are “red.” I find, surprisingly, that Illinois and Arizona (the home states of candidates Obama and McCain, respectively) did not have an “average” red/blue effect. Instead, each candidate performed better in less populous areas of their respective home state than observationally equivalent areas in other states. The linear effects indicate that outcomes varied in less populous areas; more populous areas went overwhelmingly for Obama.

While latent cultural variables may or may not break at borders, state-level policies do indeed break discretely at state lines. In the second example, I consider county-level crime rates as the dependent variable. Specifically, I look at 2006 violent and property crime rates in the United States, accounting for population, income, and poverty rates. I find a trade-off between the two types of crime. The measures of property and violent crime used are the first two factors from a battery of crimes reported at the county level, and hence are uncorrelated. My mixed penalty method find an interesting state-level trade-off: at the state level, property crime and violent crime rates correlate negatively. This, to my knowledge, is an undiscovered stylized fact.
The method casts insight into a broad array of questions beyond American voting behavior and policy outcomes. To demonstrate further the utility of the proposed method, I apply it to the G-Econ data, collected at Yale by William Nordhaus and colleagues, which estimates 1990 GDP per capita at each degree of latitude and longitude across the globe. After accounting for effects explainable by average rainfall, temperature, and average precipitation, my method finds two sets of results. First, GDP per capita is a national-level effect; I find no evidence of this value fluctuating by former colonizing power. Second, the results highlight the importance of peaceful political competition. The countries with positive state-level GDP effects either had a peaceful democratic transition (South Africa), or some form of nonviolent pluralism, at least within the ruling party (Cameroon, Kenya). The countries with negative state level effects faced violent ethnic and political conflict, either through a guerilla uprising (Mozambique) or sustained ethnic killings (Burundi).

The paper progresses in four parts. First, the method is situated within a broad class of regularization techniques. A description of smoothing and variable selection, along with my work in combining the two, follow. The three case studies are next, with a conclusion discussing insights and future directions for research.¹

2 A Brief Statistical Overview

My mixed penalty method can be viewed as a variant of several methods commonly used by political scientists. This section aims to draw out some of these conceptual links, in relation to methods both “exotic” and common. Specifically, rather than consider variable selection and smoothing as two separate classes of estimator, I present them as variants of “regularization” methods. Regularization provides a framework that unifies variable selection, smoothing, and a host of related methods. Though I take a likelihood approach in the actual modeling, this section highlights some of the conceptual connections between the method developed here and mixed-effects, empirical Bayes, and fully Bayesian specifications.

¹Simulations demonstrating the method’s efficacy, as well as R code, are available from the author upon request.
2.1 Shrinkage, bias, and regularization

Regularization can be used to frame a wide variety of statistical methods, allowing for estimates with certain desirable properties. In this case, point estimates of exactly zero for one set of variables, and “smooth” estimates for another, though the uses are far wider.\(^2\) Regularization acts to constrain the least squares or maximum likelihood (LS/ML) estimates, shrinking point estimates towards zero. The constraint is incorporated in the estimation by the researcher. This can be naturally viewed as either a prior or a constraint on the likelihood, depending on whether the parameters are assumed fixed or random; the mechanics of optimization and integration carry through identically in either case. This constraint on the likelihood (or prior) may arise through any combination of a series of concerns: reducing prediction error, maximizing the posterior, or producing consistent estimates in ill-posed models (where \(k > n\), estimates are unstable, or the log likelihood does not exist)\(^3\).

Regularization-induced shrinkage as a means of reducing prediction error has been long-recognized, through engaging in the “bias-variance trade-off.” Bradley Efron and Carl Morris (1977) provide a particularly clear explanation of the issues. Following Efron and Morris, batting averages in the first half of a baseball season can be used to predict averages in the second. The sample average of first-half estimates are unbiased estimates of second-half outcomes. These sample means, though, contain both the “true” value, and, to some extent, random variation. Those dramatically above the group mean in the first half can be expected to be a little bit closer to the group mean in the second half; shrinkage, in reducing prediction error, shrinks the estimates towards zero in order to account for regression to the mean. In fact, Efron and Morris show how shrunken estimates from the first half of the baseball season produce predictions for second-half averages notably superior to the (unbiased) sample mean in mean-squared predictive error. Recent work by Andrew Gelman and Jennifer Hill (2006) provides examples using empirical Bayes estimates that balance a within-stratum mean (the unbiased estimate) with a grand-mean (the empirical prior),

---

\(^2\)Bickel and Li (2006, with discussion) provide a rigorous, useful, and complete overview of regularization techniques across a broad array of problems.

\(^3\)Throughout, \(k > n\) refers to situations where the model contains more parameters than observations.
weighting the former towards the latter. If each stratum contains the same number of observations (is “balanced”), the shrinkage weights can be viewed as variance estimates, and ANOVA decompositions, t-statistics, and confidence intervals follow directly. Unbalanced designs have an ambiguous number of degrees of freedom, and hence can motivate a fully or empirical-Bayesian approach,\(^4\) though the fundamental insights—shrinkage towards a grand mean based on weights balancing within- and cross-group variances—carries through.

The variable selection method I employ shrinks estimates as well, but in a particularly useful manner. The method, the LASSO (*Least Absolute Shrinkage and Selection Operator*, Tibshirani 1996), shrinks estimates to either an estimate the same sign as the unconstrained estimate or zero. This produces its variable selection properties. The method, and its many extensions, have received much attention in recent applied work (see Hesterberg (2008) for a complete introduction and overview). Both fixed- and random-effects estimates are somewhere between the unbiased estimate and zero, producing a non-zero estimate for each coefficient. This results in models that can grow rather unwieldy in both presentation and interpretation. The shrinkage method I employ shrinks most parameters to *exactly* zero, leaving the remainder at some value between an unbiased estimate and zero. This produces the variable selection methods employed within.

A second motivation for regularization is to render over-identified models \((k > n)\) identifiable (Tikhonov 1943 is the seminal piece). Over-identified models have found broad applicability. The Generalized Method of Moments estimator (GMM) allows over-identified parameters to be identified, through entering a series of constraints on the estimators (“moment conditions”). The fundamental result is that these estimators, under regularity conditions, are asymptotically normal (Hansen 1982). Similarly, nonparametric smoothing methods can be viewed as estimating the projection of an infinite dimensional curve onto finite data. For the intuition of an “infinite dimensional” curve, a smooth curve that admits a Taylor expansion, with each coefficient a parameter, fits this criterion. Since \(k\) is infinite, some regularization allows for identification of the curve (Wahba 1990).

\(^4\)For those using HLMs in R, this is why the `lmer` command returns t-values but not p-values.
Bayesian simulation methods and bootstrapping follow a similar motivation. Fully identifying a posterior distribution requires the identification of an infinite number of parameters, one for each moment.\(^5\) Logistic regression with complete separation is an over-identified model, and regularization through a prior can produce sensible results (Gelman, Jakulin, Pittau, and Su 2008).

Shrinkage may also be useful in identified models. Hierarchical linear models reparameterize each within-group mean, a potentially large number relative to sample size, with a single parameter, that reflects the cross-group variance; this is the approach taken through mixed- and random-effects models. Though not commonly used, ridge regression can similarly stabilize estimates, introducing some bias, but with the payoff of reducing the standard errors on highly correlated covariates. The ridge regression objective function and mixed-effects likelihood are similarly regularized, with a constraint placed on the squares of coefficients.

The primary question in regularization is choosing the “tuning parameter,” which controls the level of shrinkage. In smoothing, this parameter controls the level of the fidelity to the data, balancing the trade between an exactly linear fit and a complete interpolation of the data. In Bayesian settings, this parameter controls the balance between the likelihood and prior. In likelihood settings, like the one I use here, the smoothing parameter is often “profiled.” This requires rewriting the tuning parameter (sometimes called a “nuisance parameter”) as a function of the coefficient estimates, and then fitting involves alternating between estimating the coefficients and then using these values to estimate the tuning parameter. This may be done through an EM type setting, as done here, or a gradient-descent method (Lindstrom and Bates 1988).

Two types of criteria can be used to select this parameter: those that try to minimize prediction error and those that approximate the posterior. When the “true” model contains an infinite number of parameters, the former is consistent, otherwise the latter is consistent (Shao 1997). Prediction based methods include cross-validation, which split the data into

---

\(^5\)Consider a distribution expanded out through its moments in its characteristic function; though the results may be “indeterminate,” in the sense that each distribution admits an expansion in this manner, but two different distributions may agree on each moment. The log-normal and the distribution \(f(x) \propto \exp -x^{\frac{1}{4}}\) are the standard example.
subsets, fit the model to all but one of the subsets, and proceed to predict the excluded data. This is done for each subset in turn, with the tuning parameter selected as the one which minimizes this value. Since cross-validated risk is an estimate, and hence random, minimizing the cross-validated risk plus one standard error is often preferred.

Two less-computationally intensive methods follow the same logic. Akaike’s Information Criterion (AIC) and Mallow’s CP estimate predictive error, which results in a penalty of $2k$ added on the estimated loss. CP is based on least-squares, while AIC is likelihood-based, though they produce the same penalty. Generalized Cross-Validation (GCV) approximates “leave-one-out” cross-validation. Its major advantage is in its computational efficiency; the estimated error from successively dropping each observation and predicting it can be estimated in a single pass of the data. Intuitively, and mathematically, the GCV criterion is close to that which minimize’s average Cook’s distance.$^6$

A second set of criteria approximate the posterior distribution. Schwarz’s Bayesian Information Criterion (BIC) approximates the model’s posterior with a normal (a Laplace approximation), and shows that, with a uniform prior over all candidate models, the log posterior is approximately the log-likelihood plus $k \log(n)$.

The method I develop in this paper is closest to this, in that I maximize a penalized likelihood. In a Bayesian framework, this corresponds with a maximum a posteriori (MAP) estimate. I impose two separate constraints. The first set of constraints smooths over geography, accounting for the correlation among nearby outcomes. The second set “selects” jurisdiction-specific variables, through producing point estimates of exactly zero for most variables.

3 Combining Smoothing and Selection

My method works through combining two different types of penalties on the likelihood in a linear model. The first component penalizes the sums of the squares of the coefficients (an

---

$^6$Specifically, replace the $h_{ii}$, the $i^{th}$ component of the hat matrix in the formula for Cook’s distance, with $h_{ii}$ and then sum over all values.
“$L_2$” penalty), which naturally “smooths” the resultant curve. The second penalizes the sums of the absolute values of the coefficients (an “$L_1$” penalty), which forces many coefficients to have point estimates of precisely zero. This section discusses the logic, rationale, and connections between these two forms of penalty.

3.1 The Smooth Component

I model the smooth component through the use of the popular nonparametric method of smoothing splines (Wahba 1990, Gu 2002, Cristianini and Shawe-Taylor 2000). Luke Keele’s (2006, 2008) recent work provides an excellent introduction to many issues on this topic. For a concise introduction to smoothing splines within a penalized regression framework, see Pearce and Wand (2006). Smoothing splines partition observed variance in the variable of interest into a penalized component and an unpenalized component. Most commonly, the linear component is unpenalized, while a “roughness penalty” is imposed on the nonlinear component.

For a broad array of functions, it can be shown that if the true $\eta$ is sufficiently smooth,\(^7\) then the population minimizer of the form $E((y - \eta)^2|X)$ can be written as $\hat{\eta} = R\hat{c} + S\hat{d}$, for $n \times 1$ vectors $c$ and $d$. $R$ is an $n \times n$ matrix purely determined by $X$ and assumptions about the nature of $\eta$, while $S$ is a low-dimensional matrix, generally linear in $X$. $R$ 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:

\[
\begin{align*}
\{\hat{c}_{SS}, \hat{d}_{SS}\} &= \arg\min_{c,d} (y - Rc - Sd)'(y - Rc - Sd) + n\lambda_2 c'R' Rc \\
&= \arg\min_{c,d} \text{var}(y - Rc - Sd) + \lambda_2 \text{var}(Rc)
\end{align*}
\]

Since $R$ is an $n \times n$ matrix, the problem has $k > n$, necessitating the regularization. The level of regularization, or, in this case, smoothing, is controlled by the parameter $\lambda_2$. For $\lambda_2 = 0$, the fitted values are a complete interpolation of the data. For $\lambda_2 \to \infty$, the fitted values approach a constant function.

\(^7\)In the case of the most commonly used splines, this requires $\int (\eta'')^2 dx < \infty$; in general, the penalty is required to be finite.
values approach the least squares line from regressing $y$ onto $S$, the unpenalized space. Selecting $\lambda_2$ controls the balance between these two extremes. The coefficients in $c$ are penalized, while those in $d$ are not.

### 3.2 The Variable Selection Component

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 prediction error, or higher posterior or penalized likelihood, as described above) 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:

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

(3)

The constraint sets some of the variables to zero, with $\lambda = 0$ giving the least squares solution and $\lambda_1 \to \infty$ returns $\hat{\beta}_{\text{LASSO}} = 0$. 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 $\beta$ and $(x)^+\text{ denote } x \cdot I(x > 0)$. In an orthogonal design,
the LASSO estimator can be written (see Tibshirani 1996: 269):

$$
\hat{\beta}_j^{LASSO} = \hat{\beta}_j^0 \left(1 - \frac{\lambda_1}{|\hat{\beta}_j^0|}\right) +
$$

(4)

The LASSO estimator shrinks least squares estimates greater than $\lambda_1$ towards zero by factor $1 - \lambda_1/|\hat{\beta}_j^0|$. Covariates with least squares estimates less than $\lambda_1$ 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. 2004). The method may also be motivated as the posterior mode, with a Laplacian prior placed over $\beta$; see Park and Casella (2008) for a fully Bayesian treatment. The primary shortcoming of a Bayesian approach is that posterior median estimates are not precisely zero, while the primary shortcoming of the frequentist approach is the difficulty in finding confidence intervals with proper coverage. I embed it here within a likelihood framework. This allows for confidence intervals on the shrunken variables, in that confidence intervals can be generated on the shrunken (penalized) values, rather than the true values. The major advantage of this approach is that it produces point estimates of zero; the loss is in only providing confidence intervals on the population parameters projected onto the constrained space.

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,\(^8\) is given below:

\(^8\) This is the constraint used in random effects models, ridge regression, or through assuming a normal prior over the coefficients. The resulting estimates differ in interpretation, based off whether $\beta$ is assumed random or fixed, but the optimization is the same.
\[
\hat{\beta}^{\text{LASSO}} = \arg \min_\beta \sum_{i=1}^{n}(y_i - x_i'\beta)^2 \text{ subject to } \sum_{j=1}^{k} |\beta_j| \leq q_{\text{LASSO}} \tag{5}
\]

\[
\hat{\beta}^{\text{smooth}} = \arg \min_\beta \sum_{i=1}^{n}(y_i - x_i'\beta)^2 \text{ subject to } \sum_{j=1}^{k} \beta_j^2 \leq q_{\text{smooth}} \tag{6}
\]

Figure 3 about here.

The geometric interpretation is made clear in figure 3. Consider the case with only two coefficients, \(\hat{\beta}_1\) and \(\hat{\beta}_2\). In this case, the ridge constraint is the circle \(\hat{\beta}_1^2 + \hat{\beta}_2^2 = k_2\). The LASSO constraint, in contrast, is the square \(|\hat{\beta}_1| + |\hat{\beta}_2| = k_1\). The confidence (Scheffe) ellipse is centered at \((\hat{\beta}_1, \hat{\beta}_2)\), and its shape is governed by \(\text{cov}(\hat{\beta}_1, \hat{\beta}_2)\). 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 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.

4 The Mixed Penalty Method: Combining Smoothing and Selection

My mixed penalty model combines the two approaches above, through simultaneously fitting a smooth curve while selecting from a set of known covariates.

4.1 The Mixed Penalty Objective Function

The target function consists of three components. The first is an unpenalized component, a matrix \(S\) with columns consisting of latitude, longitude, controls, and in intercept. The second is a matrix \(X\), composed of a set of indicator variables, by jurisdiction. Variables in this component are selected, so they are placed under a LASSO constraint. The final
component is a dense $n \times n$ matrix $R$, where the closer points $i$ and $j$, the larger the value of $r_{ij}$. This captures the geographic correlation in the data, with closer points presumed to be closer in value than points further apart. Any number of forms of $R$ can be selected; I produce each element in $R$ with a radial basis function,\(^9\) where

$$R = [r_{ij}] = \exp \left\{ - \left( (x_{i,\text{long}} - x_{j,\text{long}})^2 + (x_{i,\text{lat}} - x_{j,\text{lat}})^2 \right) \right\} \quad (7)$$

This component captures local geographic correlation. Along the main diagonal of the matrix are a series of ones, and the greater the geographic distance between two points, the lower the value in this matrix. Two points infinitely far apart are then constructed to have zero correlation. Finally, I assume that the outcome, $y$ is distributed as follows:

$$y \overset{iid}{\sim} N(X\beta + Rc + Sd, \sigma^2 I) \quad (8)$$

Adding constraints to the log-likelihood generates a mixed penalty function of the following form, with $v = \text{sign}(\beta)$:

$$\{\hat{c}_{\text{mixpen}}, \hat{d}_{\text{mixpen}}, \hat{\beta}_{\text{mixpen}}\} = \arg \min_{c,d,k,\lambda_1,\lambda_2} \left( y - Sd - Rc - X\beta \right)\left( y - Sd - Rc - X\beta \right) + n\lambda_2 c' R' Rc + q\lambda_1 v' X' X |\beta| \quad (10)$$

The two different penalties correspond with the two distinct components of change over geography, as a function of latitude and longitude. The $\lambda_2 c' R' Rc$ component puts a "smooth" constraint on how the variable shifts over the geography. The second component selects state level variables.\(^\text{10}\) The components of $S$ are unpenalized by either set of constraints.

---

\(^9\) Using thin-plate splines, a two-dimensional variant of cubic smoothing splines, resulted in practically identical results. The radial basis splines proved more stable numerically, and were easier to implement, so these were favored. Simulations revealed that setting the values of $R$ below .1 and .01 to 0 led to quicker convergence, but slightly worse performance in terms of mean squared error. The $R$ matrix is therefore kept dense throughout.

\(^\text{10}\) Note that the solutions to this minimization are not the population parameters in equation 8, but the projections of these parameters onto the constrained space. The confidence intervals provided in this paper have proper coverage on these projected variables, but not the true parameters.
Each component; unpenalized, smooth, and sparse; corresponds with a particular aspect of the functional form dictated by existing theory. The penalties allow a means of folding information about the functional form into the estimation. In less Bayesian language, the penalized likelihood introduces bias that is useful for both reducing mean-squared error and setting negligible covariates to zero.

4.2 On Variable Selection and Hypothesis Testing

The normal method for variable selection is a two-stage process. First, estimates are produced, and their precisions estimated. Second, their magnitude is judged relative to their precisions, and if it is above some threshold in absolute value, normally 1.96, they are deemed “significant.” The LASSO methodology described here attempts to combine both of those steps into one. Consider again equation 4, as an example. The method, implicitly, shrinks each estimate a certain number of standard errors, with the number of standard errors controlled by the data-driven smoothing parameter, $\lambda_1$. If the shrinkage changes the variable’s estimate’s sign, the estimate is exactly zero. This eliminates the arbitrary selection of a p-value.

LASSO-type methods produce a consistent estimate that, then given the data, asks whether each covariate would be better estimated as zero or not.\(^\text{11}\) It can be viewed, in fact, as estimating each effect and a p-value simultaneously, such that a consistent estimate is produced.

The “cost” of this procedure is a shrunken estimate, biased towards zero. A posterior distribution can be simulated (Park, Trevor, Casella & George 2008), but the posterior median or mode is never precisely zero; the Bayesian approach generates shrinkage but not selection. For that reason, I keep the method on a likelihood footing. Confidence intervals are non-symmetric, due to the constraint, but the penalized log likelihood given above allows for some sense of precision through the likelihood ratio test.

\(^\text{11}\) The asymptotic properties of LASSO estimators are sensitive to $\lim_{n \to \infty} \sqrt{n} \lambda_1$. In my method, this value approaches a constant. This leads to a conservative variable selection method, with the following properties: (1) estimates are consistent, (2) all non-zero parameters are estimated as non-zero with probability tending to one, and (3) all zero parameters are estimated as non-zero with some positive probability. In effect, the method does not “over-select” zero coefficients.
All non-zero estimates throughout the paper are significant at the standard 95% level. The bias/variance trade-off associated with LASSO estimation allows for a sharpening of coefficient estimates. The estimates that are non-zero are estimated more precisely than through normal ML/LS methods, with the zero-estimates also “soaking up” some residual variance. I only present confidence intervals for one of the models (Model 2 in table 1), as the sample size for each model is large enough for the LASSO method to automatically select a p-value less than 0.05 (as discussed above). Confidence intervals, in these models, are not intended for post-estimation variable selection. Checking whether the confidence intervals contain zero, estimate by estimate, would obviate the whole point of the procedure. The LASSO method, used here, generates consistent, shrunken point estimates, with the intent of selecting and estimating effect size, and producing parsimonious, easily interpretable models, in the same step.

4.3 The Algorithm

The algorithm is a generalized EM algorithm. The parameters $\lambda_1$ and $\lambda_2$ are taken as missing, so the “E” step consists of predicting these values, conditional on the current estimates of the other parameters. The “M” step takes a likelihood-increasing step in the negative gradient direction of the penalized likelihood. The M step takes one of two forms. A series of small steps are taken towards the spline fit, in a manner uncorrelated with the current LASSO component. Once no further spline steps are available, any feasible steps are taken in the LASSO direction, either updating current non-zero values, or adding a single additional variable. The process is iterated to convergence. Details are presented in Appendix A.

5 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. State-level boundaries are crucial to this analysis; the Electoral College apportions electors based on state
lines, and campaigns respond accordingly, targeting key states. Whether something is the matter with Kansas (Frank 2004, Bartels 2006), explaining how the effect of income on vote choice qualitatively differs between “red” and “blue” states (Gelman, Shor, Bafumi & Park 2005), or typologies of state level “cultures (Elazar 1984),” states provide a useful level of analysis. The singular importance of state lines appears every four years when Americans select a president. Popular vote is aggregated within known state lines\textsuperscript{12}, and these state-level results determine the assignation of all of the state’s electors to a single candidate. This section analyzes county-level vote returns from the 2008 presidential contest. 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.

The dependent variable throughout this section is the log-odds ratio of county-level vote returns for Barack Obama vs. John McCain. The data are from each county in the continental United States ($n = 3109$), with McCain and Obama accounting for 98.7\% of 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 logs of total number of votes cast and 2000 census population estimates correlate at .98.

The primary stylized fact I explore is America’s urban-rural divide: urban, dense areas overwhelming supported Obama, while rural, sparse areas supported McCain. I present three models of increasing complexity, each characterizing the relation between vote share and number of votes cast. The models are summarized below:

1. \textit{The basic model}: This model fits a smoothing spline over latitude and longitude, while selecting among state-level intercepts.

\textsuperscript{12}With the exception of Maine and Nebraska.
2. *The linear vote model:* This is model one, except the log of the population (2000 census estimate) is controlled for (i.e. is “unpenalized”).

3. *The saturated model:* This model extends the linear vote model by selecting among state-specific linear effects in log population, by state. The log population variable is standardized by state, i.e. de-meaned by the state-specific sample mean and divided through by the state specific standard error.

*Table 1 about here.*

Results can be found in table 1. I find three sets of results. First, rather unsurprisingly, Arizona, Oklahoma, Texas, and Utah were systematically “red states,” with California, the District of Columbia, Florida, Illinois, and Maryland as “blue.” Put more precisely, county-to-county variation between these states and their geographic neighbors is best explained with a state-specific effect (intercept), rather than simply the smooth progression of the natural geographic variation across the country.

The saturated model, though, casts some unique insight. I find two separate, distinct, patterns of partisanship across the states. There are two types of “red” or “blue” effects: states where each county relatively favors one party or the other, and states where higher population counties are more likely to favor one party or the other. This leads to a counter-intuitive interpretation: states with a *positive* intercept, but *negative* linear term favor Obama, while the opposite holds for McCain.

The reasoning behind this is clear. Obama did relatively better in more populous areas, while McCain did relatively better in less populous areas. This was captured through the unpenalized component, linear in the log of population. A negative coefficient means that counties with below-average populations, for that state, favored Obama; a positive coefficient means counties with populations below-average for that state favored McCain. The home-state effect for McCain and Obama was not an “intercept-effect,” in that it was not the case that everybody in that state preferred either candidate, on average, a bit more. Instead, the home-state effect was most noticeable in low-population areas; Obama and McCain did
better in the lower-density areas than would have been otherwise expected in their home states.

I find four states that agree in both intercept and linear trend with favoring McCain: New Jersey, Oklahoma, Texas, and Utah. Only New Jersey is a surprise, the others are deep red. The New Jersey effect is a result of the geographic nature of the analysis: relative to its neighbors in the Mid-Atlantic Coast, which are quite blue, New Jersey is relatively red. The only state that was pro-Obama in both intercept and linear trend was North Carolina, where he did better both on average and in less populous counties.

I also find two states that disagree, between sign and intercept. Each had a linear trend favoring Obama, where Obama did better in less populous areas, but an intercept effect favoring McCain: New York and Ohio. Obama won both states, though only Ohio was contested. This phenomenon seems to be explained by two factors: one large, urban area favoring Obama (New York City, Cleveland), and another favoring McCain (Buffalo, Long Island, Cincinnati), and with both geographically separated by quite a distance. Both were in pro-Obama regions (the Midwest, and Mid-Atlantic), but they both were relatively less pro-Obama than expected, after controlling for the national effect of population.

In sum, I find two different ways in which a state may be red or blue. The state’s counties may be, on average, red or blue, regardless of population; or, less populous areas may be more or less likely to favor a candidate. This builds on the key finding of Gelman, et al. (2008), that the effect of income on vote choice differs qualitatively between red and blue states. The analysis of Gelman and colleagues, though, presupposes that states, rather than regions, are useful way to stratify voters.

I find a sharp effect, confirming their assumption, in certain cases—the “redness” and “blueness” of some states breaks discretely at state lines, as does the effect of population on county-level vote returns. There are, indeed, red and blue states, and population density has differing impacts upon county-level partisanship among them.

Unfortunately, the “culture” underlying state-level partisanship is conceptually vague, and we cannot say with certainty that it “breaks” at state lines. Unlike this culture, we
do indeed know that state laws end at political borders. This patchwork of geographically
discrete laws lends itself to my mixed penalty method, as discussed next.

6 Property and Violent Crime in the United States

Criminal statutes and enforcement throughout the United States vary from state to state. For example, thirty seven states have a death penalty, forty six states have some form of hate crime law, but only thirty two of these states’ hate crime laws consider sexual orientation, and sentencing guidelines vary dramatically across states. Crimes punishable by the death penalty ranges across states, from first degree murder, to first degree murder with aggravating factors, to kidnapping and hijacking. Incarceration rates in 2006 ranged from 141 per hundred thousand, for Maine, and 835 per hundred thousand in Louisiana. Differing approaches to state-level criminal policy provides a natural variation across and within states. Using data from the 2006 Uniform Crime Reporting Program, this section looks for state level effects in county-level data across the United States.

The dependent variables are the first two factors generated from the per capita occurrence of eight types of crimes, as reported at the county level: murder, rape, robbery, assault, burglary, larceny, motor vehicle theft, and arson. The two factors account for 86% of the total variance in the eight types of crime. The first factor (viol.crime) loads heavily on murder, arson, and rape, and the second factor loads heavily on burglary, larceny, and motor vehicle theft (prop.crime). The two factors correlate at .02, so the analysis captures two different effects. Since the factors, prop.crime and viol.crime, have no natural scale, they are rescaled to range from zero to one hundred, allowing coefficients to be interpreted on a percent scale. Besides an intercept, latitude, and longitude, the percent of the population below the poverty line and median income in the county are used as controls. Similar to the previous analysis, all analyses are weighted by county population.

Table 2 contains the results from this analysis, for each of the two factors. The linear models, with only the controls and no state-level effects, explain 98% and 88% of viol.crime.

13Exploratory analysis led to converting these crimes’ occurrences to a log scale.
and prop.crime, respectively. The bulk of variation in crime is explained, as expected, by median income and poverty levels.

*Table 2 about here.*

Yet, the selected effects do reveal a stark pattern: the relative distribution between violent and non-violent crimes varies across states. States that are high in one type of crime are likely to be low in another, suggesting a tradeoff between the two types of crime. That legislators partake of some sort of “economic calculus,” weighing costs and benefits (Becker 1968), is not clear, but the results suggest that the relation between incarceration, apprehension, and prevention can be influenced at the state level.

The reason for this variance is not obvious upon further analysis. Comparing the state-level effects to states that have the death penalty, number of people executed, incarceration rates, changes in incarceration rates, and expenditures on corrections budgets revealed no clear patterns. Correlations with Elazar’s state-level cultures, as in Fisher and Pratt (2006), again revealed no pattern. The literature finds that

“..there is no lack of arguments for the primacy of a single factor, such as incarceration, the economy, or abortion. But, generalizing across all sectors, the crime drop does appear to derive from some beneficial synergies, mutual reinforcements among the factors we have discussed (Blumstein and Wallman 2006: 143).”

The clearest finding generated by my method is that there appears to be a state-level tradeoff between violent and property crimes, which, to my knowledge, is an undiscovered phenomenon.

7 The Geographic Distribution of GDP across Africa

The G-Econ project, undertaken by William Nordhaus and colleagues, has produced a dataset which estimates GDP at each degree of latitude and longitude across the globe, as well as a battery of geographic and climatological covariates.\footnote{Details and complete documentation available at http://www.gecon.yale.edu. Accessed March 9, 2009.} 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 dummy variables by country; this misses the natural correlations due to geographic proximity between-and within-countries. It does not fully account for within-country variation, cross-country variation due to the smooth progression of geographic outcomes across borders, and 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, Satyanath & Sergenti 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 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 three models with a geographic component, trying to identify sustained mean-level effects corresponding by 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. Variables were weighted by 1990 population.

The following three specifications are presented:

1. The basic models: This model fits a smoothing spline over latitude and longitude, con-
trolling for log population, elevation, temperature, and rainfall. State-level intercepts are selected.

2. *The former colonizer model*: This is model one, except a new set of indicators for former colonizing power are included.

3. *The saturated colonizer model*: This is model two, except the indicators for colonizing power are also interacted with log population, and then selected using the LASSO.

*Table 3 about here.*

Results for the three models are in table 3. I find evidence of cross-national variation in the effect of population on GDP; the linear terms that were prevalent in the US county-level partisanship data are not present here. Each selected variable is a country-level intercept; none of the former-colonizer indicator variables were selected. Country-to-country variation in GDP per capita seems best conceived as a country-specific effect, rather than clustering by former colonizer. Though this is clearly not evidence that these former-colonizer effects are absent, they are small in relation to country-specific effects.

I only analyze the results that are robust across the three models, which eliminates many states that were identified in Model 1 but not selected in the later models. The most striking reading of these findings highlight the important relation between civil society and economic productivity. Though the very concept “civil society” is fraught with ambiguity and controversy (e.g. Hassan 2009), the results from this analysis stand out. The data, from 1990, capture Africa at the tail end of the “third wave” of democratization (Huntington 1991), though an analysis of Polity III data suggested that Africa was further lagged behind the rest of the world in this trend (Jaggers & Gurr 1995).

In 1990, Cameroon, 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 & 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.

Burundi and Mozambique stand in stark contrast to Cameroon, Kenya, and South Africa. In 1990, Burundi was fraught with ethnic conflict, with political divides corresponding with a ethnic difference between the minority Tutsi, who effectively control the government, and the majority Hutu. This longstanding divide, compounded with a range of institutional failures, devolved into widespread violence and accounts for Burundi’s lack of productivity (Ndikumana 1998). 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.
8 Conclusion

Empirical analysis entails a clear tension between the desire to fit a complex model while also producing simple results. The method here was designed with the hopes of both striking at a fundamental problem in the study of government, while producing tractable and simply interpreted models. While geography certainly matters, incorporating this fact into a statistical analysis is difficult. We often utilize spatial models to explain rational choice, and yet the physical space within which political behavior actually occurs is the hardest to incorporate.

This paper began with a consideration of the role geography and political boundaries play in our social world. The data structure, smooth over a two dimensional space with discontinuous jumps at known locations, is, to my knowledge, unique to political science. The statistical problems proved challenging, yet tractable. The state-line/sub-state data structure required some nonparametric model, as the different categories (states) were geographically correlated, with the covariance structure among them unknown. Existing analyses reach to a series of dummy variables based on proximity to a river or port. The coefficient on a dummy variable gives a mean effect, but it may not everywhere be the same mean. The spline component addresses this shortcoming. Similarly, incorporating fixed or random effects into smoothers works well, but suffers from the problem of producing complex, hard to present models. In the models presented here, between 1/2 and 19/20 of the covariates under the LASSO constraint were set to zero—a boon for both researchers looking to present results, and readers looking to understand them.

The substantive insights would not have been discernible within more complex models, as the researcher often struggles to determine which results are noise and which ones are not. Type I error always lurks around the corner, and arbitrary p-values do not help. Instead, I am able to make some solid claims. In the United States, several states are red and several are blue. While population was key in predicting a county’s percent vote for Obama versus McCain, in several states this effect varied from the national average. The nature of the
variation indicated that urban areas went almost uniformly for Obama, while Obama and McCain made inroads against each other in rural areas. In terms of state level crime rates, I find a negative correlation between property-crime and violent-crime. This suggests, at either the legislative or cultural level, some tradeoff between these two types of crimes.

Regarding GDP per capita in Africa, my first result is dispositive: country-to-country variation in GDP per capita, after controlling for basic controls and smooth geographic fluctuation, is best considered a nation-level phenomenon, rather than a phenomenon that clusters by former colonizer. The state level effects that are selected highlight rather clearly the relationship between a government that allows a political space where disagreements can be handled peacefully, and those where ethnic conflict and insurgency occur.

In conclusion, this paper was intended to both introduce a political science audience to a class of methods unified by the idea of regularization (Bayesian, penalized likelihood, and others), linking it to concerns of model selection, variable selection, and appropriate criteria. Incorporating these concerns into our statistical analyses, in a manner driven by data structure suggested by theory, can lend insight to deep problems otherwise unapproachable.
A The Algorithm

A.1 A brief overview of algorithmic concerns

Two classes of algorithms may be used to optimize this particular loss function: gradient descent algorithms and regularization path algorithms. Gradient descent algorithms take steps in the negative gradient direction, perhaps augmented by exact or approximate local curvature information. Regularization path approaches require (i) characterizing the coefficients as a function of the tuning parameters \( (\lambda_1, \lambda_2) \)\(^{15}\) and (ii) an external criterion such as GCV, CP, AIC, or BIC for evaluating the fit. Methods may trace the entire regularization path, such as the Least Angle Regression algorithm of Bradley Efron and colleagues (2004), or they may start at a point and move along the regularization in a manner which increases the criterion function (see Wright, et al. for an example).

My method combines the two approaches. The primary issue is that the regularization path of the spline is continuous, while the regularization path of the LASSO component is piecewise-linear. My algorithm integrates the two. Smooth steps are taken, while avoiding intrusion into the current space spanned by the LASSO covariates. When no further spline steps can be taken, a LASSO step is taken. The process is iterated to convergence.

A.2 The algorithm

The EM algorithm used in this model consists of two different steps. The “E” step involves predicting the smoothing parameters, conditional on current coefficient and residual estimates. The “M” step consists of three separate steps. First, the ideal step in the spline direction is calculated, taking into account covariance with the space spanned by the active set. Second, if a spline step that can reduce the residual sums of squares is possible, it is taken. If not, the LASSO estimates in the active set are updated. If no feasible step is possible in either the spline space or LASSO space, the algorithm ends.

\[
\hat{\beta}_{LASSO}^-, \hat{c}^- = 0 \quad \text{and} \quad \lambda_1^- = \lambda_2^- = 0.
\]

Generate matrix \( X \), each column a dummy

\(^{15}\)The function traced out by the coefficients while moving the tuning parameters through their range is the “regularization path.”
variable; matrix $R$ parameterizing the smooth component; and matrix $S$ containing the intercept, latitude, longitude, and any other unpenalized covariates for each observation.

Define the following, with $v = \text{sign} (\beta)$:

$$L(\beta, c, d, \lambda_1, \lambda_2) := (y - Sd - Rc - X\beta)'(y - Sd - Rc - X\beta) + \lambda_2 ||Rc||_2^2 + \lambda_1 v'X'X|\beta|$$

Define the subdifferentials of $L$ with respect to $d$, $\beta$, and $c$, where the subscript denotes the penalty, “0” for least squares, “1” for the LASSO penalty, and “2” for the smoothing spline penalty, and $\epsilon$ the current estimate of the residuals:

$$\frac{\partial L}{\partial d} = \nabla_0 = -2S'\epsilon \quad (11)$$

$$\frac{\partial L}{\partial \beta} = \nabla_1 = -2X'\epsilon + 2q\lambda_1 X'Xv \quad (12)$$

$$\frac{\partial L}{\partial c} = \nabla_2 = -2R'\epsilon + 2n\lambda_2 R'Rc \quad (13)$$

The algorithm progresses as follows:

1. The “E” Step: Given current estimates of the smoothing parameters. This is done by first setting each subdifferential in equations 12-13 to zero, taking the variance of both sides, and then solving for each estimate.

$$E(\lambda_1|\hat{\epsilon}, \hat{\beta}, \hat{c}, \hat{d}) = \left( \frac{\hat{\epsilon}'XX'\hat{\epsilon}}{q\hat{v}'XX'X\hat{v}} \right)^{1/2} \quad (14)$$

$$E(\lambda_2|\hat{\epsilon}, \hat{\beta}, \hat{c}, \hat{d}) = \left( \frac{\hat{\epsilon}'RR'\hat{\epsilon}}{\hat{\epsilon}'R'R\hat{R}'R\hat{c}} \right)^{1/2} \quad (15)$$

2. The “M” step: Calculate the step in the gradient descent direction that maximizes the likelihood. This requires maximizing over step-length $\eta_0$, $\eta_1$, and $\eta_2$ the following:
\[
\{\hat{\eta}_0, \hat{\eta}_1, \hat{\eta}_2\} = \arg \max_{\{\eta_0, \eta_1, \eta_2\}} \mathcal{L}(\beta - \eta_1 \nabla_1, c - \eta_2 \nabla_2, d - \eta_0 \nabla_0, \lambda_1, \lambda_2) \tag{17}
\]

Taking partials with respect to \(\eta_0, \eta_1, \) and \(\eta_2\) gives the following set of equations:

\[
\begin{bmatrix}
\nabla'_0 S' S \nabla_0 & \nabla'_1 X' S \nabla_0 & \nabla'_2 R' S \nabla_0 \\
\nabla'_0 S' X \nabla_1 & \nabla'_1 X' X \nabla_1 & \nabla'_2 X' R \nabla_2 \\
\nabla'_0 S' R \nabla_2 & \nabla'_1 X' R \nabla_2 & \nabla'_2 R' R \nabla_2 \\
\end{bmatrix}
\begin{bmatrix}
\eta_0 \\
\eta_1 \\
\eta_2 \\
\end{bmatrix}
= 
\begin{bmatrix}
\epsilon' S \nabla_0 \\
\epsilon' X \nabla_1 + q' \lambda_1 \nabla'_1 X' X v \\
\epsilon' R \nabla_2 + n' \lambda_2 \epsilon' R' R \nabla_2 \\
\end{bmatrix}
\tag{18}
\]

\(\eta_0, \eta_1,\) and \(\eta_2\) can be solved for directly. If the new value of \(\eta_1\) changes a sign for \(v\), or if it is negative, the process is iterated. Note how the off-diagonal entries in the matrix above forces step-length in the spline direction to take into account the LASSO space.

Update \(c\) and \(d\).

If this reduces the residual sum of squares by more than some suitably small number (i.e., \(10^{-8}\%\), but so long as this number is small, the algorithm will work), update \(c\) and \(d\). Leave \(\beta\) unchanged.

If it does not reduce the residual sum of squares by more than the lower threshold, use the LARS algorithm (Efron, et al. 2002) to update each coefficient in the active set.

3. If there are no feasible LASSO steps, end.
Figure 1: The systematic component of the curve used in simulations.
Figure 2: 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.
Figure 3: The difference between the $L_1$ (LASSO) and $L_2$ constraints, illustrated graphically. The LASSO constraint will naturally set variables equal to zero, while the ridge constraint will not.
<table>
<thead>
<tr>
<th>State</th>
<th>Model 1 Intercept</th>
<th>Model 2 Intercept</th>
<th>Conf Interval, Model 2</th>
<th>Model 3 Intercept</th>
<th>Linear Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alabama</td>
<td>-0.54099</td>
<td>0</td>
<td>(-0.45547, .81489)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Arizona</td>
<td>-0.01596</td>
<td>-0.48252</td>
<td>(-0.48254, -0.14330)</td>
<td>0</td>
<td>4.31283</td>
</tr>
<tr>
<td>California</td>
<td>1.48004</td>
<td>0.46413</td>
<td>(0.46142, 2.31270)</td>
<td>5.49121</td>
<td>0</td>
</tr>
<tr>
<td>Connecticut</td>
<td>0.04822</td>
<td>0</td>
<td>(-0.00130, 0.029172)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>District of Columbia</td>
<td>0.96790</td>
<td>1.00389</td>
<td>(1.0038, 2.55102)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Florida</td>
<td>0.38243</td>
<td>0.14827</td>
<td>(0.14287, 0.57460)</td>
<td>0</td>
<td>-1.67614</td>
</tr>
<tr>
<td>Georgia</td>
<td>0</td>
<td>0</td>
<td>(-0.00000, 1.16429)</td>
<td>0</td>
<td>-1.04412</td>
</tr>
<tr>
<td>Illinois</td>
<td>0.63161</td>
<td>0</td>
<td>(-0.00000, 1.78043)</td>
<td>0</td>
<td>-0.35931</td>
</tr>
<tr>
<td>Kansas</td>
<td>0</td>
<td>0</td>
<td>(-1.53713, 0.00000)</td>
<td>0</td>
<td>1.14917</td>
</tr>
<tr>
<td>Maryland</td>
<td>0.62188</td>
<td>0.43734</td>
<td>(0.437284, 0.50837)</td>
<td>0</td>
<td>-6.02932</td>
</tr>
<tr>
<td>Massachusetts</td>
<td>0.24995</td>
<td>0</td>
<td>(-0.00000, 0.89383)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>New Jersey</td>
<td>0</td>
<td>-0.01967</td>
<td>(-0.32686, -0.01966)</td>
<td>-0.27447</td>
<td>0.53073</td>
</tr>
<tr>
<td>New Mexico</td>
<td>0</td>
<td>0</td>
<td>(-0.00000, 3.10682)</td>
<td>3.21603</td>
<td>0</td>
</tr>
<tr>
<td>New York</td>
<td>0.81993</td>
<td>0</td>
<td>(-0.28969, 0.00000)</td>
<td>-0.54992</td>
<td>-1.34497</td>
</tr>
<tr>
<td>North Carolina</td>
<td>0</td>
<td>0</td>
<td>(-0.00000, 1.20291)</td>
<td>0.34875</td>
<td>-0.30915</td>
</tr>
<tr>
<td>Ohio</td>
<td>0</td>
<td>-0.12864</td>
<td>(-1.08384, -0.12863)</td>
<td>-2.97038</td>
<td>-0.06309</td>
</tr>
<tr>
<td>Oklahoma</td>
<td>-0.89936</td>
<td>0</td>
<td>(-2.09429, 0.00000)</td>
<td>-4.56292</td>
<td>1.68942</td>
</tr>
<tr>
<td>Tennessee</td>
<td>-0.38062</td>
<td>0</td>
<td>(-0.92313, 0.00000)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Texas</td>
<td>-0.22442</td>
<td>-0.30812</td>
<td>(-0.64401, -0.30811)</td>
<td>-2.17515</td>
<td>0.55835</td>
</tr>
<tr>
<td>Utah</td>
<td>-0.77757</td>
<td>-0.73121</td>
<td>(-2.44162, -0.731207)</td>
<td>-11.22877</td>
<td>0.48260</td>
</tr>
</tbody>
</table>

Table 1: Results for state level analyses. Positive coefficients indicate pro-Obama tendency; negative coefficients indicate a pro-McCain effect. Results are on a log-odds scale.
<table>
<thead>
<tr>
<th>State</th>
<th>Violent Crime</th>
<th>Property Crime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alabama</td>
<td>−6.06</td>
<td>0</td>
</tr>
<tr>
<td>California</td>
<td>−15.93</td>
<td>0</td>
</tr>
<tr>
<td>Colorado</td>
<td>6.61</td>
<td>0</td>
</tr>
<tr>
<td>Florida</td>
<td>9.68</td>
<td>−38.84</td>
</tr>
<tr>
<td>Georgia</td>
<td>5.37</td>
<td>0</td>
</tr>
<tr>
<td>Illinois</td>
<td>−339.16</td>
<td>199.81</td>
</tr>
<tr>
<td>Kansas</td>
<td>4.57</td>
<td>0</td>
</tr>
<tr>
<td>Kentucky</td>
<td>−12.09</td>
<td>15.66</td>
</tr>
<tr>
<td>Maine</td>
<td>3.85</td>
<td>0</td>
</tr>
<tr>
<td>Maryland</td>
<td>0.73</td>
<td>0</td>
</tr>
<tr>
<td>Massachusetts</td>
<td>4.70</td>
<td>−0.69</td>
</tr>
<tr>
<td>Michigan</td>
<td>20.42</td>
<td>0</td>
</tr>
<tr>
<td>Minnesota</td>
<td>21.44</td>
<td>0</td>
</tr>
<tr>
<td>Mississippi</td>
<td>−31.68</td>
<td>36.48</td>
</tr>
<tr>
<td>Missouri</td>
<td>11.08</td>
<td>0</td>
</tr>
<tr>
<td>New York</td>
<td>0</td>
<td>−86.08</td>
</tr>
<tr>
<td>North Carolina</td>
<td>28.58</td>
<td>0</td>
</tr>
<tr>
<td>Ohio</td>
<td>14.24</td>
<td>0</td>
</tr>
<tr>
<td>Oregon</td>
<td>11.96</td>
<td>−4.86</td>
</tr>
<tr>
<td>Pennsylvania</td>
<td>0</td>
<td>−7.83</td>
</tr>
<tr>
<td>South Carolina</td>
<td>7.01</td>
<td>0</td>
</tr>
<tr>
<td>South Dakota</td>
<td>−13.46</td>
<td>12.69</td>
</tr>
<tr>
<td>Tennessee</td>
<td>16.19</td>
<td>0</td>
</tr>
<tr>
<td>Texas</td>
<td>41.48</td>
<td>0</td>
</tr>
<tr>
<td>Utah</td>
<td>7.52</td>
<td>−2.14</td>
</tr>
<tr>
<td>Virginia</td>
<td>0</td>
<td>40.09</td>
</tr>
<tr>
<td>Washington</td>
<td>33.84</td>
<td>−24.87</td>
</tr>
<tr>
<td>West Virginia</td>
<td>0</td>
<td>8.30</td>
</tr>
<tr>
<td>Wisconsin</td>
<td>13.31</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2: State level effects on property and violent crime.
<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Intercept</td>
<td>Intercept</td>
<td>Intercept</td>
</tr>
<tr>
<td>Algeria</td>
<td>0.04513</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Burundi</td>
<td>-0.02952</td>
<td>-0.00681</td>
<td>-0.00682</td>
</tr>
<tr>
<td>Cameroon</td>
<td>0.14262</td>
<td>0.09752</td>
<td>0.09770</td>
</tr>
<tr>
<td>Democratic Republic of Congo</td>
<td>-0.08895</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ethiopia</td>
<td>-0.05443</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ghana</td>
<td>-0.02596</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Kenya</td>
<td>0.08671</td>
<td>0.07477</td>
<td>0.07490</td>
</tr>
<tr>
<td>Libya</td>
<td>0.08454</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Malawi</td>
<td>-0.04017</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Morocco</td>
<td>-0.08028</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mozambique</td>
<td>-0.25903</td>
<td>-0.28682</td>
<td>-0.28734</td>
</tr>
<tr>
<td>Nigeria</td>
<td>0.03882</td>
<td>0.04967</td>
<td>0.04976</td>
</tr>
<tr>
<td>Rwanda</td>
<td>-0.01722</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Senegal</td>
<td>-0.04065</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>South Africa</td>
<td>0.13345</td>
<td>0.12786</td>
<td>0.12809</td>
</tr>
<tr>
<td>Uganda</td>
<td>-0.01213</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Zimbabwe</td>
<td>0.01626</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3: Results for analyses on Africa. Positive coefficients indicate higher GDP per capita. Results are on a log-odds scale.
References


C Supplemental Materials: Simulation Evidence

C.1 Simulation Set-up

Before handling the observed data, I present a series of simulations that mimic a smooth, two-dimensional curve with a few “jumps.” A 5x5 grid is drawn over the range \( \{0, 10\} \times \{0, 10\} \). Samples ranging between 400 and 2,500 are drawn uniformly across this range. The systematic smooth and jump components of the target function are:

\[
\eta_{\text{smooth}}(x, y) = \sin\left(\frac{(x + y)^2 \cdot \pi}{500}\right) + \sin\left(\frac{(x - y)^2 \cdot \pi}{200}\right) \\
\eta_{\text{jump}}(x, y) = I(2 < x < 4, 2 < y < 4) - 2 \cdot I(6 < x < 8, 4 < y < 6)
\]  

(19) (20)

The target function was shown earlier, in figure 1. It is comprised of a smooth curve laid over a 5x5 grid, with one square sticking up and on dropped down above two squares. Since the possible location of these breaks are known, the \( X \) matrix has 25 columns and \( n \) rows, with each column a dummy variable. Points are selected uniformly from across the square, and independent, normal noise is added to generate the data. The goal is to fit a smooth curve over the smooth component while simultaneously selecting the jumps. The following six simulations are run:

\[
\begin{align*}
\text{Model Specifications:} \\
\text{sim}_{\text{jump}}(x_i, y_i) &= \eta_{\text{smooth}}(x_i, y_i) + \eta_{\text{jump}}(x_i, y_i) + u_i \\
\text{sim}_{\text{nojump}}(x_i) &= \eta_{\text{smooth}}(x_i, y_i) + u_i
\end{align*}
\]

Variance Specification:

\( u_i \sim iid N(0, 1) \)

Sample Sizes Used:

\( n \in \{400, 1225, 2500\} \)
Simulations were run 300 times, with the same set of data fit to a thin plate smoothing spline, a thin plate smoothing spline with fixed effects by state,\textsuperscript{16} and my mixed penalty algorithm.

C.2 Simulation Results

The simulations demonstrate the mixed penalty method’s utility. Table C.2 shows the root mean square (RMS) difference between the estimated and true curve. My mixed penalty method uniformly outperforms the spline, both in the presence and absence of actual discontinuities. In simulations with discontinuities, the mixed penalty method performs better than the misspecified spline model, which is as expected. It also performed comparably to a fixed-effects specification, even though this specification uses more degrees of freedom and produces many more point estimates. When there are indeed no discontinuities, the method performs nearly identically to a smoothing spline, which is the “true” model, in this case. The fixed-effects method overfits in these scenarios, while the mixed penalty avoids this. All three models perform well, in that no model grossly out-performs the others. The mixed penalty method performs comparably in terms of RMS error, but also has the added advantage of reasonably accurate variable selection. Very few false positives are made, while the method is acceptably powerful, selecting large effects almost every time, and small effects with a frequency increasing in sample size.

\textit{Table C.2 and Figure 4 about here.}

Figure 4 illustrates the selection properties, showing the distribution of the estimates of the coefficients that are in truth not zero. While the true effect is either 1 or -2, both the mixed penalty method and fixed effects grossly underestimate its magnitude. The spline component naturally mutes any discontinuities. The mixed penalty method produces sharper estimates of a higher magnitude, on average, than the fixed effects, for the larger effect. In forcing many covariates to zero, the method is able to go “all-in” on effects that it does

\textsuperscript{16}Splines are fit using function \texttt{ssanova} in R library \texttt{gss}, using a REML estimate of the smoothing parameter. Fixed effects specifications include columns of the $X$ matrix entered as partial splines.
Figure 5 shows the distribution of estimates for coefficients that are in truth zero. This demonstrates the real payoff to the method: estimates that are zero, or suitably close, are estimated at zero, with “suitably close” being determined by the data. A large number of “noisy” covariates that model complex models are reduced to zero. Table 5 reenforces this point. In models with jumps, non-zero terms are selected between half and three-fourths of the time, with decreasing standard error. The larger effect was selected almost every time, so the smaller effect goes from being selected almost never \((n = 400)\) to selected about half the time \((n = 2500)\). Those correctly estimated as zero goes down, but the effect size is quite small, especially relative to fixed effects. In models with no jumps, those correctly selected as zero are all above 95%; the model can reduce to a smoothing spline when appropriate. The mixed penalty method shrinks an appreciable number of estimates to exactly zero, while the remainder are both centered on zero and much smaller in magnitude than their fixed effects counterparts.

The simulations demonstrate the utility of the method on reasonable sample sizes, ranging from hundreds to thousands of observations. My mixed penalty method can select both nonzero and zero estimates with reasonable accuracy, and, when there are in truth no jumps, the method outperforms smoothing splines. For large effect size, the estimates of non-zero effects are larger and sharper than the fixed effects specification. The method presents a sound way to fit a model with both a smooth component and a small number of jumps.
Table 4: Root mean square difference between true curve and fitted curve times 1000, by whether jumps are present or not, sample size, and method used for fitting.

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Spline</th>
<th>Spline+Fixed Effects</th>
<th>Mixed Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>n=400</td>
<td>29.32</td>
<td>26.92</td>
<td>27.25</td>
</tr>
<tr>
<td>n=1225</td>
<td>16.67</td>
<td>15.53</td>
<td>15.60</td>
</tr>
<tr>
<td>n=2500</td>
<td>11.71</td>
<td>10.97</td>
<td>10.90</td>
</tr>
<tr>
<td>Mean</td>
<td>19.23</td>
<td>17.81</td>
<td>17.92</td>
</tr>
<tr>
<td>n=400</td>
<td>24.39</td>
<td>23.45</td>
<td>24.39</td>
</tr>
<tr>
<td>n=1225</td>
<td>14.00</td>
<td>13.50</td>
<td>13.97</td>
</tr>
<tr>
<td>n=2500</td>
<td>9.78</td>
<td>9.44</td>
<td>9.67</td>
</tr>
<tr>
<td>Mean</td>
<td>16.05</td>
<td>15.46</td>
<td>16.01</td>
</tr>
</tbody>
</table>
Figure 4: Distribution of Non-Zero Coefficient Estimates.
Figure 5: Distribution of Zero Coefficient Estimates.
<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter Value</th>
<th>n=400</th>
<th>n=1225</th>
<th>n=2500</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correctly Selected as Non-Zero</td>
<td>0.5067</td>
<td>0.5833</td>
<td>0.7250</td>
</tr>
<tr>
<td></td>
<td>Mean Estimate, Mixed Penalty</td>
<td>0.1938</td>
<td>0.1867</td>
<td>0.1893</td>
</tr>
<tr>
<td></td>
<td>Standard Error, Mixed Penalty</td>
<td>0.1324</td>
<td>0.0598</td>
<td>0.0594</td>
</tr>
<tr>
<td></td>
<td>Mean Estimate, Fixed Effects</td>
<td>0.2021</td>
<td>0.1757</td>
<td>0.1678</td>
</tr>
<tr>
<td></td>
<td>Standard Error, Fixed Effects</td>
<td>0.1792</td>
<td>0.0933</td>
<td>0.0643</td>
</tr>
<tr>
<td><em>With Discontinuities</em></td>
<td>Correctly Selected as Zero</td>
<td>0.9430</td>
<td>0.8693</td>
<td>0.7511</td>
</tr>
<tr>
<td></td>
<td>Mean Estimate, Mixed Penalty</td>
<td>0.0001</td>
<td>0.0004</td>
<td>0.0007</td>
</tr>
<tr>
<td></td>
<td>Mean Abs Value, Mixed Penalty</td>
<td>0.0051</td>
<td>0.0070</td>
<td>0.0108</td>
</tr>
<tr>
<td></td>
<td>Standard Error, Mixed Penalty</td>
<td>0.0245</td>
<td>0.0233</td>
<td>0.0262</td>
</tr>
<tr>
<td></td>
<td>Mean Estimate, Fixed Effects</td>
<td>0.0403</td>
<td>0.0358</td>
<td>0.0391</td>
</tr>
<tr>
<td></td>
<td>Mean Abs Value, Fixed Effects</td>
<td>0.0474</td>
<td>0.0495</td>
<td>0.0468</td>
</tr>
<tr>
<td></td>
<td>Standard Error, Fixed Effects</td>
<td>0.0974</td>
<td>0.0647</td>
<td>0.0542</td>
</tr>
<tr>
<td><em>Without Discontinuities</em></td>
<td>Correctly Selected as Zero</td>
<td>0.9848</td>
<td>0.9700</td>
<td>0.9634</td>
</tr>
<tr>
<td></td>
<td>Mean Estimate, Mixed Penalty</td>
<td>−0.0009</td>
<td>−0.0018</td>
<td>−0.0028</td>
</tr>
<tr>
<td></td>
<td>Mean Abs Value, Mixed Penalty</td>
<td>0.0022</td>
<td>0.0026</td>
<td>0.0024</td>
</tr>
<tr>
<td></td>
<td>Standard Error, Mixed Penalty</td>
<td>0.0191</td>
<td>0.0162</td>
<td>0.0133</td>
</tr>
<tr>
<td></td>
<td>Mean Estimate, Fixed Effects</td>
<td>0.0538</td>
<td>0.0430</td>
<td>0.0445</td>
</tr>
<tr>
<td></td>
<td>Mean Abs Value, Fixed Effects</td>
<td>0.1046</td>
<td>0.0698</td>
<td>0.0584</td>
</tr>
<tr>
<td></td>
<td>Standard Error, Fixed Effects</td>
<td>0.1136</td>
<td>0.0617</td>
<td>0.0410</td>
</tr>
</tbody>
</table>

Table 5: Simulation results for estimates that are in truth zero, and those that are not.