Bootstrapped Basis Regression with Variable Selection: A New Method for Flexible Functional Form Estimation

Brenton Kenkel* Curtis S. Signorino†

January 23, 2013

Work in progress: comments welcome.

Abstract

We introduce a new statistical model to estimate the shape of potentially nonlinear, multivariate relationships. The method is built on basis regression, in which simple functions like polynomials are combined to approximate more complex relationships. To reduce the instability typically associated with basis regression, we use penalized regression techniques that perform automatic model selection, eliminating many terms from the final estimate. We focus on methods that satisfy the oracle property, which guarantees that terms with no true effect on the outcome are excluded from the estimated model in sufficiently large samples. Finally, we calculate standard errors and other estimates of variability via the bootstrap. In a series of simulations, we show that our method can accurately estimate nonlinear relationships, even if the exact functional form is not known in advance. We apply our method to Gartzke's (2007) data on the "capitalist peace" and find that joint democracy and trade dependence may increase the chance of conflict in some cases, contrary to the original results.

*Ph.D. candidate, Department of Political Science, University of Rochester (email: brenton.kenkel@gmail.com)

†Associate Professor, Department of Political Science, University of Rochester (email: curt.signorino@rochester.edu)

1 Introduction

Most empirical research in political science uses simple statistical models to capture complex social processes. These models impose specific functional forms on the relationship between the covariates and the outcome, even though political scientists can rarely claim to know the exact form the relationship should take. Not all statistical models require such restrictions: it is possible to directly estimate the functional form of a relationship, rather than imposing one in advance. This paper encourages political scientists to use flexible modeling and provides a new technique for doing so. This method, bootstrapped basis regression with variable selection, is tailored to be accessible to political scientists and easy to implement in typical empirical applications, while maintaining the flexibility and good statistical properties of other nonparametric techniques. We also provide software, the `polywog` package for R, to implement this technique in a user-friendly way.

At the core of our method is *basis regression*, which involves representing a complex functional form as a linear combination of simpler functions (Efromovich 1999; Eubank 1999). The simplest example of this procedure is polynomial regression, in which the dependent variable is regressed on a polynomial expansion of the covariates. In other words, the right-hand side of the regression equation contains the covariates, higher-order powers of them, and interactions among them. As more terms are included in the basis expansion, the model becomes more flexible, allowing it to approximate more complex underlying functions. The estimated coefficients represent the weight to be placed on each term in the approximation, meaning the results cannot typically be interpreted by inspection of individual coefficients. Instead, as with quadratic and interaction terms in ordinary regression models, graphical methods are crucial for interpreting the results of basis regression. This paper provides multiple examples of how to use plots for interpretation, and our `polywog` software provides a user-friendly implementation of predicted-value plots.

The next part of our method is *variable selection* via penalized regression. Depending on the number of covariates, basis regression may entail regressing the outcome variable on hundreds or even thousands of terms. Ordinarily, this means the estimates have large standard errors, making inference difficult. We mitigate this problem by using penalized regression techniques that automatically perform model selection, yielding a coefficient of zero for most terms in order to obtain more precise estimates of the remaining coefficients. The procedures we use are conceptually

similar to stepwise regression, but have better statistical properties and are more computationally efficient. Specifically, we use variable selection techniques that have the *oracle property*, such as the adaptive LASSO (Zou 2006) and the smoothly clipped absolute deviations (SCAD) estimator (Fan and Li 2001). Under the oracle property, the probability of an irrelevant term being included in the estimated model goes to zero as the sample size increases. By using variable selection techniques with the oracle property, our method prevents against overfitting, allowing for the estimated model to be as complex or as parsimonious as the data warrants. For example, if the true data-generating process is linear, then each higher-order terms of the basis expansion is unlikely to be included in the estimated model, provided that the sample is sufficiently large. Therefore, although our method is flexible enough to approximate complex nonlinear functions, it is also capable of yielding results similar to a simpler model—when the true model is simple.

The final step in our method is the *bootstrap*, which we use to calculate variability in the estimates. Researchers are typically interested in functions of the results, such as predicted probabilities or marginal effects, rather than the raw estimates. Bootstrap simulation makes it easy to compute standard errors and confidence intervals for these substantively interesting quantities. In addition, from a technical standpoint, bootstrap simulation helps avoid problems that arise in the analytic computation of standard errors from penalized regression models.

We are by no means the first to encourage wider adoption of nonparametric modeling in political science. Beck and Jackman (1998) advocate the use of generalized additive models, which are more flexible than traditional linear models but typically allow for few or no interactive relationships (see Hastie and Tibshirani 1990). Such models are also examined in Keele's (2008) general review of flexible regression techniques for social scientists. Beck, King and Zeng (2000) encourage political scientists to use neural nets, a popular technique from artificial intelligence and machine learning. Spline-based methods for testing hypotheses about the shape of relationships are introduced by Wand (2011, 2012). Most recently, Hainmueller and Hazlett (2012) develop a new nonparametric estimator, kernel-regularized least squares. We believe all of the methods mentioned here should be part of a political scientist's toolkit; there is no single technique that is best for every individual application. The main advantage of our method over these alternatives is the oracle property, which ensures that the chance of irrelevant terms being included in the estimated model decreases with sample size. This property allows researchers with large datasets to be confident that nonlinearity

in the estimated functional form represents a real feature of the data-generating process, rather than overfitting to noise in the sample. Our method is thus well suited for researchers who want to guard against functional form misspecification but nonetheless suspect that the true form of the relationship in their data is linear.

The remainder of the paper proceeds as follows. In Section 2, we discuss why functional form misspecification is an important problem and illustrate the general philosophy of nonparametric regression. We explain our proposed technique in formal detail in Section 3. Section 4, presents the results of a series of Monte Carlo experiments we ran to assess the finite-sample performance of the technique across various types of data. In Section 5, we replicate the empirical analysis of Gartzke (2007) using our technique. We find much more variance in the effects of joint democracy and economic interdependence than the original analysis allowed for; in many cases, these variables increase the chance of conflict. Concluding remarks are given in Section 6.

2 Functional Form Misspecification and Nonparametric Models

In political science, it is rare for a researcher to have strong *a priori* knowledge of the exact functional form of the relationship among variables. Even formal models usually are not so closely tailored to observed data that they suggest specific functional relationships to be included in regression specifications.[1] There is thus an element of guesswork in setting up an appropriate regression model for a research problem in political science. The presence of this uncertainty should propel political scientists to use methods that are robust to specification error, so as to ensure that their results are not driven by unwarranted functional form assumptions. Unfortunately, such techniques are still uncommon in political science. The dominant approach to statistical modeling is to use a regression model with a restrictive specification, limiting the forms that the estimated relationships may take. In these models, the bulk of the covariates are simply included linearly and thus assumed to have a constant effect on the outcome. Some flexibility may be added by including higher-order powers or interaction terms, which allow a variable's effect to depend on its own value or that of another term. However, such terms usually are only included for the covariates of primary interest (as opposed to "control variables"), and only when a nonlinear or interactive relationship is explic-

[1] Of course, there are some exceptions; see the examples in Wand (2011, 2012). For more on the relationship between formal models and nonparametric estimation, see Kenkel and Signorino (2012*b*).

itly suggested by one of the main hypotheses. These restrictive models carry a risk for researchers: they may yield biased results and inappropriate inferences if the functional form assumptions are not satisfied.

To illustrate the perils of the standard approach, consider a researcher who wants to estimate the relationship between the variable of interest X and the outcome Y, while controlling for Z. She hypothesizes that, holding Z constant, an increase in X will be associated with an increase in Y. To test this, she estimates a standard linear regression model,

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon, \tag{1}$$

and assesses whether the estimated β_1 is statistically discernible from 0. She may then also use the coefficient estimates to gauge the magnitude of the change in Y that would result from a change in X and judge whether the relationship is substantively significant. Each element of this inferential process depends on equation (1) being the correct specification of the model. Suppose, on the contrary, that the true relationship is nonlinear. For example, the data-generating process may be represented by an S-curve, as in the following equation:

$$Y = \Phi(X + Z) + \epsilon. \tag{2}$$

If this is the true model, then the results of an ordinary regression model will be misleading. By construction, the model cannot capture the fact that the marginal effect of X on the outcome depends on its own value and that of Z, approaching no effect at all when $|X + Z|$ is large. As such, when the researcher estimates substantive effects from the coefficients of a linear model, they will show the effect of a one-unit increase in X to be the same as when $(X, Z) = (0, 0)$ as when $(X, Z) = (3, 3)$. Moreover, this kind of misspecification causes bias in the estimated standard errors of $\hat{\beta}$, which in turn means the nominal significance level of any hypothesis tests will be misstated. The coefficient estimates in any given sample may lead the researcher to conclude correctly that X has a positive relationship with Y, but beyond that the results of a linear model are misleading about the nature of the relationship.

Even comparatively mild forms of functional form misspecification can also threaten the validity

of inferences from standard regression models. Consider the following model, in which the variable of interest indeed has a linear relationship with Y, but the control variable enters the equation nonlinearly:

$$Y = \beta_0 + \beta_1 X + f(Z) + \epsilon, \tag{3}$$

with f being some nonlinear function of Z. Unlike in the previous example, it is theoretically possible for linear regression to yield an accurate representation of the relationship between X and Y (though it cannot do the same for the "control variable" Z). Specifically, the estimate of β_1 must be unbiased, or at least consistent—which will not be the case if X is correlated with $f(Z)$. The functional form misspecification in this case can actually be thought of as a type of omitted variable bias, in which the relevant term $f(Z)$ is omitted from the estimated model.[2] Therefore, the usual condition for unbiasedness, which is that the included variables be uncorrelated with the omitted ones, applies in this case. If it is not met, we run into some of the same problems as in the previous example: the nominal significance level of the hypothesis test that $\beta_1 > 0$ will be incorrect, possibly leading to overconfidence, and estimates of the substantive relationship between X and Y will be biased. For these reasons, functional form misspecification cannot be ignored even if it only affects variables not of primary interest to the researcher.

Luckily for applied analysts, there are ways to obtain consistent estimates of the relationship between covariates and the outcome without knowing its exact functional form. Such techniques fall under the rubric of nonparametric regression. In traditional (parametric) regression models, a particular functional form is assumed, such as the linear model in equation (1), and its coefficients are estimated. Nonparametric regression entails directly estimating the relationship without imposing any particular form in advance. To continue the example from above, a nonparametric model would simply assume

$$Y = \mu(X, Z) + \epsilon, \tag{4}$$

where μ is a function that maps from two-dimensional space onto the real line. The estimation procedure uses the data to form an estimate of μ. Consistency in this context means that, as the sample size grows large, the distribution of estimated $\hat{\mu}$ approaches a point mass on the true

[2] For more detail on functional form misspecification as a form of omitted variable bias, see Kmenta (1986, pp. 449–450) and Signorino and Yilmaz (2003).

relationship. If the true relationship is linear, then the estimated functions will be approximately linear in large samples. On the other hand, if the true model is nonlinear, the nonparametric technique will be able to capture this as well. Nonparametric regression can be used either in place of a traditional regression model, or alongside one. In the latter case, the nonparametric model would be used as a robustness check to demonstrate that the traditional results do not differ much from the flexible estimate, indicating that the functional form restrictions made in the original model are warranted.

The main practical difference between parametric and nonparametric regression is in the interpretation of the results. Nonparametric methods directly estimate the functional form of μ, so the effect of any particular variable cannot be reduced to a single coefficient. The simplest method is to plot the estimated relationship between one or two covariates and the outcome, along with confidence intervals computed via bootstrap simulation. This general technique is already familiar to political scientists, who often use predicted-value plots for substantive interpretation of logistic regression and other generalized linear models (King, Tomz and Wittenberg 2000; Hanmer and Kalkan 2012). It is also useful to compute marginal effects, which are partial derivatives of the predicted outcome $\hat{\mu}(X, Z)$ with respect to X or Z. A researcher can calculate the marginal effect at various observations of (X, Z) in order to determine if the effects are essentially constant, in which case a linear model might be warranted, or if they vary across observations. A summary measure somewhat analogous to the regression coefficient on X would be the average marginal effect, which is the mean of the estimated marginal effect $\partial\hat{\mu}(X, Z)/\partial X$ across each value of (X, Z) in the dataset.

Nonparametric regression has two main drawbacks: computational cost and statistical efficiency. The first of these is relatively minor. It is true that nonparametric techniques require more calculations than ordinary linear regression, but advances in processing power have made this less of a concern. Computation times vary widely across models (and the particular implementations of each one), but in general it is now feasible to apply nonparametric methods to datasets of typical size for political science. The second drawback, statistical efficiency, is perhaps more serious. The flexibility of these models entails using more degrees of freedom than ordinary linear models do, which in turn means there is more variation in the estimates. However, most nonparametric regression models allow researchers to control the flexibility of the model—and thus the amount of

additional variability—via a set of tuning parameters. These parameters govern how much the estimated function "smooths" the raw data, between the extremes of a simple linear model (maximal smoothing) and fitting a line through each data point (minimal smoothing). Users of nonparametric methods usually select the tuning parameters via cross-validation, which favors models that generalize well to out-of-sample data.[3] Therefore, even with medium-sized datasets, researchers can balance flexibility with precision when using nonparametric models.

3 Method

We now turn to discussion of our new method for flexible functional form estimation. Conceptually, our proposed method for functional form estimation is simple and consists of three elements:

1. A **simulation** loop applied to
    2. A **variable selection** method applied to
        3. A **basis regression** model.

There are a number of specific techniques that may be employed for any of the above elements. We address these in reverse order.

3.1 Basis Regression

The goal of regression analysis is to estimate the conditional expectation of a response variable Y, given a vector of covariates $X = (X_1, \ldots, X_p)$. We assume the researcher possesses N observations $(x_1, y_1), \ldots, (x_N, y_N)$ with generic element (x_i, y_i). The regression problem can be represented by the equation

$$y_i = \mu(x_i) + \epsilon_i, \tag{5}$$

where μ represents the (unknown) conditional expectation function and ϵ_i is white noise error. In parametric regression analysis, the function μ is assumed to be known up to a finite set of parameters. For example, the functional form $\mu(x_i) = x_i'\beta$ yields the ordinary linear regression model,

$$y_i = x_i'\beta + \epsilon_i, \tag{6}$$

[3] For more detail on the use of cross-validation to select tuning parameters, see Section 3.2 below.

for which the task is to estimate β. No such functional form restrictions are made in nonparametric analysis. Instead, nonparametric methods attempt to estimate the functional form of μ directly, only making comparatively weak assumptions like continuity or differentiability of the conditional expectation function. What makes nonparametric estimation difficult is that the function μ lies within an infinite-dimensional space, while any given collection of data has only finitely many degrees of freedom. It is nonetheless possible to obtain consistent and asymptotically normal estimates of μ, albeit with slower convergence rates than in parametric models (see Pagan and Ullah 1999).

Our technique of choice is series estimation, or basis regression, which consists of approximating μ via a linear combination of basis vectors (see Andrews 1991; Newey 1994; Efromovich 1999).[4] The key assumption is that there exist a sequence of basis functions $\{h_m\}_{m=1}^{\infty}$ and a sequence of scalar weights $\{c_m\}_{m=1}^{\infty}$ such that

$$\mu = \sum_{m=1}^{\infty} c_m h_m. \tag{7}$$

This condition is easily met in practice if μ is assumed to be continuous. The Weierstrass Approximation Theorem states that any continuous function with a bounded domain can be approximated arbitrarily well by a sequence of polynomials (Eubank 1999, p. 122). Therefore, as long as μ is continuous, the condition can be satisfied with a power series basis. In multi-index notation,[5] this is equivalent to setting

$$h_m(X) = X^{\alpha_m},$$

where each α_m is a distinct p-vector of non-negative integers. For example, with $p = 2$, the first few terms of the sequence would be[6]

$$\{h_m(X)\} = \{1, X_1, X_2, X_1^2, X_1X_2, X_2^2, \ldots\}.$$

Basis functions other than power series include fixed-knot splines (Newey 1994), trigonometric

[4] There are many other families of nonparametric regression methods, too numerous to review in detail here. Pagan and Ullah (1999) provide a general overview of nonparametric estimation, largely with a focus on kernel methods. Non-series methods are reviewed briefly in Efromovich (1999, ch. 8). Hastie, Tibshirani and Friedman (2009) review numerous methods for flexible estimation of unknown relationships, albeit from a machine-learning standpoint rather than a strictly statistical one.

[5] Multi-index notation entails $X^{\alpha} \equiv \prod_{j=1}^{p} X_j^{\alpha_j}$ for vectors $X, \alpha \in \Re^p$ (see Gallant 1981).

[6] When discussing power series, we adopt the convention that the degree of the m'th term, $\sum_{j=1}^{p} \alpha_{mj}$, is non-decreasing with m.

series (Eubank 1999, ch. 3.4), and Fourier series (Gallant 1981).

The basis expansion conveniently allows us to use standard parametric techniques to obtain a nonparametric estimate of μ. In effect, it transforms the problem into a simple linear regression. By taking the basis expansion up to the M'th term, the regression equation (5) can be rewritten as

$$y_i = \sum_{m=1}^{M} c_m h_m(x_i) + \eta_i, \tag{8}$$

where the residual term η_i represents

$$\eta_i = \sum_{m=M+1}^{\infty} c_m h_m(x_i) + \epsilon_i. \tag{9}$$

Equation (8) is simply a linear model. The regressors are the basis expansion $(h_1(x_i), \ldots, h_M(x_i))$, the coefficients are the weights $(c_1, \ldots, c_M)$, and the error term is η_i. This suggests that the unknown function μ can be estimated by ordinary least squares regression. To derive the estimator, we first must introduce some notation. Let $\mathbf{y} = (y_1, \ldots, y_N)$ denote the vector of observed responses, and let $\mathbf{X}$ denote the $N \times p$ matrix with i'th row x_i. In addition, it will be convenient to denote $h^M(X) = (h_1(X), \ldots, h_M(X))$ for $X \in \Re^p$, and to let $H^M(\mathbf{X})$ be an $N \times M$ matrix with i'th row $h^M(x_i)$. We may now write the M'th-order series estimator of $\mu(X)$ as $\hat{\mu}^M(X)$,[7] which is defined as

$$\begin{aligned} \hat{\mu}^M(X) &= \sum_{m=1}^{M} \hat{c}_m^M h_m(X) \\ &= h^M(X)' \hat{c}^M \end{aligned} \tag{10}$$

where the weights are estimated via

$$\hat{c}^M = [H^M(\mathbf{X})' H^M(\mathbf{X})]^{-1} H^M(\mathbf{X})' \mathbf{y}. \tag{11}$$

Andrews (1991) establishes asymptotic properties for this type of estimator. He finds broad regularity conditions under which series estimators are consistent and asymptotically normal, provided that $M \to \infty$ as $N \to \infty$ at an appropriate rate. In practice, M is usually chosen via a data-

[7] We drop the superscript and simply write $\hat{\mu}(X)$ in contexts where this will not result in any confusion.

	$d=1$	$d=2$	$d=3$	$d=4$	$d=5$
$p=2$	3	6	10	15	21
$p=3$	4	10	20	35	56
$p=5$	6	21	56	126	252
$p=10$	11	66	286	1001	3003
$p=15$	16	136	816	3876	15504

Table 1. Number of terms contained in a power series expansion of degree d for a model with p covariates.

dependent procedure like cross-validation. When $\{h_m\}$ is a power series, it is natural to select M so that all terms up to some degree d are included. We discuss the selection of d in greater detail in the next subsection.

Basis regression can also be used to estimate functions of μ, including the marginal effects of covariates. The nonparametric model allows for the regressors to have a non-linear or interactive relationship with the outcome, meaning that the marginal effects of $\mu(X)$ may vary with X. To compute the M'th-order series estimator of $\partial\mu/\partial X_m$, we can simply take the derivative of $\hat{\mu}^M$,

$$\frac{\partial \hat{\mu}^M(X)}{\partial X_m} = \sum_{m=1}^{M} \hat{c}_m^M \frac{\partial h_m(X)}{\partial X_m}, \tag{12}$$

where the weights are computed using (11). The average marginal effect of X_m across the distribution of X can then be estimated via

$$\frac{1}{N}\sum_{i=1}^{N} \frac{\partial \hat{\mu}^M(x_i)}{\partial X_m}. \tag{13}$$

Series estimation of derivatives and other functions of μ is discussed at further length in Andrews (1991).

3.2 Variable Selection

Like any other statistical technique, basis regression is subject to the bias-variance tradeoff. Each additional basis term included in the regression makes the model more flexible, allowing the estimated conditional expectation to approximate more complex functions, but also increases the variability of the estimate. Moreover, when there are more than a couple of covariates, extra flexibility may require the inclusion of a vast number of basis terms. To illustrate this, Table 1 shows

the number of terms in a power series expansion up to degree d of p covariates.[8] For example, with just ten covariates—a moderate amount by the standards of empirical political science—a third-degree power series expansion requires regressing $\mathbf{y}$ on 286 terms. Even with a reasonably large sample, the resulting estimates are liable to be unstable.

To preserve flexibility without inducing too much variability, we employ regularization techniques from the machine learning literature. These methods augment the least-squares criterion with a penalty on the total magnitude of the estimated coefficients (other than the intercept), which acts to keep the estimates from reacting wildly to small variations in the data. We focus on regularized estimators that also perform variable selection, by which we mean that many (often most) of the coefficients are estimated as exactly 0, so the corresponding terms can be thought of as excluded from the model. This procedure may sound similar to stepwise regression—the model selection technique best known among political scientists, mainly as an object of derision (King 1986, p. 669). The problems with stepwise regression are well known (e.g., Hurvich and Tsai 1990; Derksen and Keselman 1992), and accordingly we eschew its use. Instead, we apply more recently developed methods that are less computationally burdensome and have more attractive statistical properties.

Our regularization method of choice is the adaptive LASSO (Zou 2006), which extends the LASSO procedure originally proposed by Tibshirani (1996). We will first explain the adaptive LASSO and its properties with reference to a standard regression setup, then extend its application to basis regression. Regression of $\mathbf{y}$ on $\mathbf{X}$ via the adaptive LASSO proceeds in two steps. First, one obtains a consistent initial estimate of the regression coefficients $\hat{\beta}$, typically via OLS.[9] The

[8] The general formula for the number of power series terms required is

$$\sum_{\delta=0}^{d} \binom{p+\delta-1}{p-1}.$$

Each term in the summation represents the number of distinct multi-indices of degree δ, which is equivalent to the number of selections of size δ from a set of p elements with repetition.

[9] For simplicity of notation, we introduce the adaptive LASSO only in terms of a standard linear model with a continuous response. However, it is conceptually simple to extend the technique to other settings, such as generalized linear models where the response variable is binary or a count. In these cases, the first term of equation (14) is replaced with the relevant loss function, such as the log-likelihood for generalized linear models, and the initial estimates used as weights should be an minimizer of that function. We examine the logistic regression case in Section 4.2.

adaptive LASSO estimate of β is obtained in the second step as the solution to

$$\underset{\beta \in \Re^p}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (y_i - x_i'\beta)^2 + \lambda \sum_{j=2}^{p} \frac{|\beta_j|}{|\hat{\beta}_j|} \right\}. \tag{14}$$

The first coefficient, β_1, is presumed to be an intercept and thus is not penalized. Each subsequent term is penalized by a factor inversely proportional to its initial coefficient, so terms with initial estimates close to 0 are more likely to be excluded from the estimated model. The term λ controls the overall level of penalization. If $\lambda = 0$, the estimator is the same as OLS; greater values of λ imply more aggressive model selection, with fewer terms included in the estimated model. We discuss procedures for the selection of λ below.

The most important feature of the adaptive LASSO is the oracle property (see Fan and Li 2001). Suppose the true relationship between X and Y is given by the standard regression equation (6), and moreover that X contains a number of "irrelevant" variables. That is, there exists a non-empty set $\mathcal{J} = \{j \,|\, \beta_j = 0\}$ containing the indices of the variables that have no effect on Y. The oracle property has two components:

1. As the sample size increases, the probability of selecting the correct model goes to 1. Formally, for a sequence of estimators $\{\hat{\beta}_N\}$, $\lim_{N\to\infty} \Pr(\{j \,|\, \hat{\beta}_{Nj} = 0\} = \mathcal{J}) = 1$.

2. The asymptotic distribution of the remaining (non-zero) components of $\hat{\beta}$ is the same as that of the "oracle model" in which Y is regressed only on the relevant variables.

These properties are obviously attractive, as they suggest that the estimator may perform as well as the oracle model in large samples, without requiring *a priori* knowledge of which variables are irrelevant. Another prominent example of an estimator with the oracle property is regression with a smoothly clipped absolute deviation penalty, or SCAD, developed by Fan and Li (2001). Unlike the adaptive LASSO, SCAD does not require a consistent first-stage estimate of the coefficients, which is beneficial when there are more covariates than observations.

The oracle property can be thought of as a stronger form of the ordinary consistency property. An estimator $\hat{\beta}$ is consistent if it converges in probability to the true parameter β, which merely requires that the coefficients on the irrelevant terms approach zero in the limit. However, this may be the case even if these terms are never estimated as *exactly* zero in finite samples, as is the case

with OLS. Although OLS is consistent, it almost never yields estimates of exactly zero for irrelevant terms. The oracle property requires more from an estimator. As the sample size increases, not only must the coefficients on irrelevant terms approach zero, but the probability of their being exactly zero must approach one—a property known as consistent model selection. This requires that the estimator be sparse, estimating some terms as exactly zero even in finite samples, unlike ordinary estimators like OLS, which almost always yield non-zero estimates for all parameters. The oracle property also requires that consistent model selection not come at the expense of inefficiency; the estimator must have the same asymptotic variance as the oracle model (from which the irrelevant terms are excluded in advance). Ordinary consistency has no such requirement, so estimators like OLS are consistent even though they may have much higher asymptotic variance than the oracle model. Before moving on to the extension of the adaptive LASSO to basis regression, it is worth noting that the benefits of the oracle property do not come without a cost. In particular, oracle estimators may exhibit poor finite-sample behavior, and their risk, or maximal mean squared error across the parameter space, exceeds that of maximum likelihood estimators (Leeb and Potscher 2008).[10] We perform simulations to assess the finite-sample behavior of the adaptive LASSO and SCAD as applied to basis regression in Section 4.

It is simple to extend the application of the adaptive LASSO to basis regression. The first-stage estimates should be the $\hat{c}^M$ from ordinary basis regression, computed via equation (11), and the design matrix used in the second stage should be $H^M(\mathbf{X})$. The resulting estimates will themselves be basis weights, which can be plugged into (10) to form an estimate of the function $\mu(X)$. In the context of basis regression, the oracle property is particularly important when $\{h_m\}$ is a power series. In this case, an expansion of degree $d \geq 2$ will contain each variable in $\mathbf{X}$, the corresponding quadratic terms, and each two-way interaction. Regression specifications in political science are usually restricted to a subset of these terms: the main effects, along with a small number of quadratic terms or interactions. If this kind of restriction is warranted, meaning the true data-generating process only depends on these first- and second-order terms, then the oracle property ensures that the correct model will be selected with high probability in sufficiently large samples. On the other hand, if the true relationship is too complex to be captured in a standard specification, our method will yield more accurate results than if a restrictive model were imposed

[10] We are grateful to Marc Ratkovic for drawing our attention to this issue.

a priori. In short, our technique allows for a parsimonious estimate when parsimony is warranted, and a more complex estimate when it is not.

Up to this point, we have treated the number of basis vectors and the LASSO penalty factor λ as fixed. In practice, these should be selected in a data-driven way so that the flexibility of the model increases appropriately with sample size. The most common way of selecting these tuning parameters is via K-fold cross-validation (Hastie, Tibshirani and Friedman 2009, pp. 241–245). This is a means of approximating the out-of-sample error of the estimator under many possible configurations of tuning parameters without having to collect additional data. The first step is to select a grid of candidate values for λ and M (or the degree d, in the special case where a polynomial basis is used). Each observation in the dataset is then assigned to one of the K folds. For each candidate pair (λ, M), the model is fit K times, each time leaving out one of the folds. The out-of-sample error is then calculated via

$$\text{CVE}(\lambda, M) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_{(i)}(\lambda, M))^2, \tag{15}$$

where $\hat{y}_{(i)}(\lambda, M)$ is the fitted value from the model that excluded the fold containing i. The final model is fit on the full dataset using the tuning parameters that cross-validation finds to minimize the out-of-sample error.

3.3 Bootstrap Simulation

To make inferences from the results, we need an estimate of the variability in $\hat{\mu}$. Zou (2006) and Fan and Li (2001) provide standard error formulas for the adaptive LASSO and SCAD respectively, but these only apply to the terms with non-zero estimated coefficients. But the standard errors of $\hat{\mu}$ depend on the variability in all of the basis weights, even those estimated as zero in a particular sample. Our technique of choice is the bootstrap, which entails re-estimating the model on B synthetic datasets constructed using the original data. The nonparametric (pairwise) bootstrap or a residual bootstrap can be used, although the latter is not available for logistic models. In either case, the result is a $B \times M$ matrix of bootstrap estimated weights, which can be used to compute standard errors and confidence intervals of the basis weights $\hat{c}^M$, the conditional expectation $\mu(X)$, the average marginal effects, and other quantities of interest (Efron and Tibshirani 1993, ch. 13–

14). We typically re-estimate λ via K-fold cross-validation within each bootstrap iteration, as this produces more conservative estimates than re-using the initial λ (Tibshirani 1996), albeit at higher computational cost. However, we have found that the optimal degree of a power series basis is typically stable, and thus hold this fixed. For derivations of the asymptotic properties of the residual bootstrap for the adaptive LASSO, see Chatterjee and Lahiri (2011).

The bootstrap results also allow us to construct an alternative estimate of μ. The bootstrap-aggregated, or bagged, estimate of μ is formed by averaging the estimates computed in each bootstrap iteration (Breiman 1996). Formally, the bagged estimate of μ is given by

$$\hat{\mu}^M(X) = \frac{1}{B}\sum_{b=1}^{B} h^M(X)'\hat{c}^{M[b]}, \tag{16}$$

where $\hat{c}^{M[b]}$ represents the estimated basis weights from the b'th bootstrap iteration. Bootstrap aggregation is normally used to increase accuracy in unbiased but highly variable nonparametric techniques such as random forests. Our primary means of variance reduction is the adaptive LASSO penalty, so bagging would appear to be of less use here. Nonetheless, we have found that a bagged estimate of $\mu(X)$ are sometimes preferable to the ordinary estimate, particularly for values of X far from the mean $E[X]$.[11]

4 Monte Carlo Simulations

In this section, we use Monte Carlo simulations to examine the finite-sample performance of our estimator. There are three sets of simulations. In the first, the data-generating process is a polynomial function, meaning the true model is a subset of the power series basis expansion and thus the oracle property applies. The second setup is almost exactly the same, except the outcome variable is binary rather than continuous. Finally, in the third simulation, we examine the performance of the estimator when applied to a non-polynomial model that can only be approximated by a power series basis. In each case, we find that the penalized basis regression estimator usually outperforms both a naive linear model and unpenalized basis regression in terms of out-of-sample prediction error. Moreover, we find that our estimator is not much less efficient than the oracle model when

[11] Details available on request.

the sample size is sufficiently large.

4.1 Polynomial Equation, Continuous Response

In the first Monte Carlo experiment, the conditional expectation is a polynomial function of the covariates, as summarized in the following equations.

$$\begin{aligned}\mu(X) = 1 + X_1 - X_2 + X_3 - X_4 + X_5 - X_6 + X_1X_2 - X_3X_4 + X_5X_6 \\ - 0.25\, X_1^2X_2 + 0.25\, X_3X_4^2 - 0.25\, X_5^2X_6,\end{aligned} \tag{17}$$

$$y_i = \mu(x_i) + \epsilon_i. \tag{18}$$

In addition to the six relevant covariates, there are two irrelevant variables X_7 and X_8. A degree-3 power series expansion of X therefore consists of 165 terms, including the constant, of which just 13 enter the true model. All eight regressors are distributed normally, $X \sim N(0, \Sigma)$, where Σ contains 1 along the diagonal and ρ in each off-diagonal entry. We examine correlation values $\rho \in \{0, 0.25, 0.5, 0.75\}$. The error term is also normally distributed, with variance set at $\sigma^2 = 10$ to be roughly equal to that of Y in the case where $\rho = 0$. The number of observations generated in each iteration is $N \in \{200, 500, 1000, 5000\}$. We run $T = 500$ Monte Carlo iterations for each combination of ρ and N.

Our main interest is in the validity of the out-of-sample fitted values of various estimators, including the penalized basis regression techniques introduced in Section 3. We focus on each estimator's *mean integrated squared error*, or MISE, which is the squared error of the fitted values $\hat{\mu}(X)$ averaged over the distribution of X. To approximate the MISE of an estimator $\hat{\mu}$, we first collect the estimates calculated across the T Monte Carlo iterations, $(\hat{\mu}_1, \ldots, \hat{\mu}_T)$. We then draw $N_o = 10{,}000$ new values $(x_1, \ldots, x_{N_o})$ from the population distribution of X and compute the MISE as

$$\text{MISE}(\hat{\mu}) \simeq \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_o}\sum_{i=1}^{N_o}(\hat{\mu}_t(x_i) - \mu(x_i))^2. \tag{19}$$

This is similar to the quantity that we minimize when selecting tuning parameters in cross-validation, given above in equation (15). Lower values of the MISE are preferable, as they represent less out-of-sample error.

N	ρ	ALASSO	SCAD	OLS	naive	oracle	null
200	0.00	5.98 (0.47)	3.74 (0.67)	211.59 (−17.90)	4.08 (0.64)	0.76 (0.93)	11.20
200	0.25	5.80 (0.33)	3.79 (0.56)	213.12 (−23.63)	3.96 (0.54)	0.78 (0.91)	8.65
200	0.50	5.31 (0.21)	3.88 (0.42)	212.15 (−30.47)	3.96 (0.41)	0.83 (0.88)	6.74
200	0.75	3.52 (0.27)	2.93 (0.39)	209.83 (−42.42)	3.66 (0.24)	0.82 (0.83)	4.83
500	0.00	1.50 (0.87)	0.74 (0.93)	8.64 (0.22)	3.65 (0.67)	0.29 (0.97)	11.13
500	0.25	1.64 (0.81)	0.80 (0.91)	8.87 (−0.03)	3.53 (0.59)	0.28 (0.97)	8.60
500	0.50	1.72 (0.74)	0.92 (0.86)	8.81 (−0.32)	3.52 (0.47)	0.28 (0.96)	6.69
500	0.75	1.83 (0.62)	1.06 (0.78)	8.82 (−0.84)	3.28 (0.32)	0.29 (0.94)	4.79
1000	0.00	0.51 (0.95)	0.34 (0.97)	2.68 (0.76)	3.51 (0.68)	0.14 (0.99)	11.11
1000	0.25	0.54 (0.94)	0.33 (0.96)	2.68 (0.69)	3.39 (0.60)	0.14 (0.98)	8.58
1000	0.50	0.61 (0.91)	0.36 (0.95)	2.73 (0.59)	3.39 (0.49)	0.14 (0.98)	6.67
1000	0.75	0.74 (0.85)	0.42 (0.91)	2.71 (0.43)	3.14 (0.34)	0.14 (0.97)	4.78
5000	0.00	0.08 (0.99)	0.03 (1.00)	0.36 (0.97)	3.40 (0.69)	0.03 (1.00)	11.09
5000	0.25	0.07 (0.99)	0.03 (1.00)	0.36 (0.96)	3.28 (0.62)	0.03 (1.00)	8.57
5000	0.50	0.07 (0.99)	0.03 (1.00)	0.37 (0.94)	3.28 (0.51)	0.03 (1.00)	6.66
5000	0.75	0.09 (0.98)	0.07 (0.99)	0.37 (0.92)	3.03 (0.36)	0.03 (0.99)	4.77

Table 2. Results of the first Monte Carlo simulation. Each entry represents the approximate mean integrated squared error (MISE) of each estimator for a given sample size N and correlation ρ, computed using 10,000 out-of-sample draws from the distribution of X. The values in parentheses represent the proportional reduction in squared error compared to the null model, $1 - \text{MISE}_{method}/\text{MISE}_{null}$.

We calculate the MISE for six estimators of μ. The first three are basis regression models: one penalized via the adaptive LASSO, one penalized via SCAD, and one computed via OLS with no penalization.[12] A power series basis of degree 3 is used in all cases, and the penalization tuning parameter for the adaptive LASSO and SCAD is selected via 10-fold cross-validation in each iteration. The next estimator we consider is a "naive OLS" of $\mathbf{y}$ on $\mathbf{X}$, leaving out all interactions and higher-order terms. The last two estimators represent best- and worst-case benchmarks respectively. The best case is the oracle model, which entails regressing $\mathbf{y}$ only on the terms that appear in equation (17). Recall that the oracle property of the adaptive LASSO and SCAD implies that their sampling distributions, and hence their MISE, should converge to that of the oracle model as the sample size grows large. The worst case estimator is the null model, which ignores covariate information and always predicts the sample mean, $\hat{\mu}(X) = \bar{y}$.

The results of this set of simulations are summarized in Table 2. The cells in each row give the calculated MISE of each estimator of μ for the given combination of N and ρ. The entry in

[12] All penalized basis regression estimates are implemented via our own R package `polywog` (Kenkel and Signorino 2013), which calls functions from `glmnet` to implement the LASSO (Friedman, Hastie and Tibshirani 2010) and uses `ncvreg` to implement SCAD (Breheny and Huang 2011).

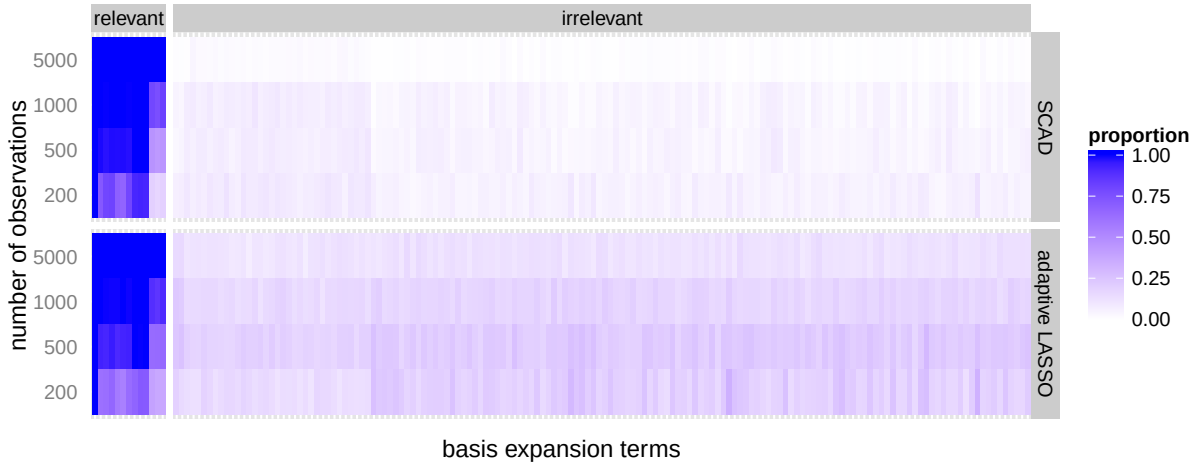

Figure 1. Heat map to illustrate the oracle property in the results of the first Monte Carlo simulation. Each cell represents a term included in the polynomial basis expansion, shaded according to the proportion of times it is included in the estimated model for the given number of observations. The small box on the left contains the 13 relevant terms, while the larger box on the right contains the 152 terms not included in the true model.

parentheses in each cell represents the proportional reduction in average out-of-sample error for the given estimator compared to the null model. Taken as a whole, the results suggest that the oracle-penalized estimators have few drawbacks relative to either ordinary basis regression or the usual linear model, regardless of sample size.[13] The naive linear estimator's MISE is comparable to that of the penalized estimators only for the smallest sample size examined. However, the naive model's performance barely improves as N increases, whereas the penalized estimators quickly become more accurate. The difference between the penalized and unpenalized basis regression models is even starker: for all values of N and ρ, the penalized estimators have a substantially lower MISE. In fact, in most cases with $N \leq 500$, the MISE of the unpenalized model is even higher than that of the null model. This may at first seem contradictory, since the null model is our purported worst case estimator. This is resolved by noting that regression models are only guaranteed to have better *in-sample* mean squared error than a null model. The MISE measures out-of-sample performance, at which an unstable estimator like unpenalized basis regression, which uses 165 degrees of freedom,

[13] It is apparent from Table 2 that SCAD outperforms the adaptive LASSO in each case. This is specific to the particular data-generating process used here, not a general feature of the two penalization techniques. In similar simulations, we have found that either technique may perform better, depending on the particular form of μ. There appears to be no systematic pattern in data-generating processes for which the adaptive LASSO is preferred over SCAD, or vice versa. However, both penalized techniques always outperform unpenalized basis regression. Details are available on request.

may be worse than the sample mean, which only uses 1.

The simulation results also show the oracle property at work in the adaptive LASSO and SCAD models. The first key feature of the oracle property, consistent model selection, is illustrated in Figure 1. Each cell in the heat map represents one of the 165 terms included in the basis expansion and is shaded by the proportion of times it has a non-zero coefficient in the estimated model. Under both the adaptive LASSO and SCAD, all of the relevant terms are almost always included in the estimated model when $N = 5000$. The probability of irrelevant terms being included decreases steadily with the sample size; the SCAD estimator does particularly well in this respect, with a very small probability of including irrelevant terms when N is large. The other important feature of the oracle property is that the penalized model must not be asymptotically inefficient relative to the oracle model. This is apparent in Table 2, which shows that the MISE of the adaptive LASSO and SCAD estimators appears to converge to that of the oracle model as the sample size grows large. At $N = 5000$, both of the penalized estimators offer very little loss in efficiency compared to the oracle model.

4.2 Polynomial Equation, Binary Response

In the second Monte Carlo simulation, the basic data-generating process is almost the same as in the first, but the response variable is now binary. As in standard binary modeling contexts, the probability of the response is modeled as a logistic transformation of an underlying index function. We use the polynomial in equation (17) as this index function. The model can be formally represented with the following equations:

$$y_i^* = \mu(x_i) + \epsilon_i, \tag{20}$$

$$y_i = \begin{cases} 1 & y_i^* \geq 0, \\ 0 & y_i^* < 0, \end{cases} \tag{21}$$

where $\mu(X)$ is given by equation (17) and ϵ_i has a logistic distribution. The scale parameter for the distribution of ϵ_i is set so that its variance is 10, the same as before. The distribution of the covariates X is also the same as in the previous simulation. We run $T = 500$ iterations for each combination of $\rho \in \{0, 0.25, 0.5, 0.75\}$ and $N \in \{500, 1000, 5000\}$.

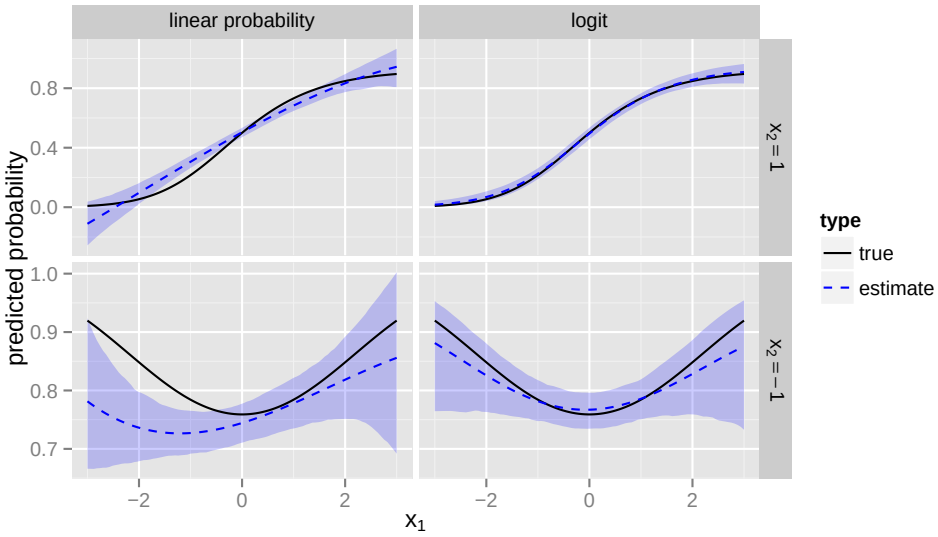

Figure 2. Illustration of fitted values in the second Monte Carlo simulation. The two columns represent the linear model and logistic model respectively, both penalized via the adaptive LASSO (using linear model weights), in the set of trials with $N = 5000$ and $\rho = 0.25$. Points in the graph represent the predicted probability of $Y = 1$ for the given value of X_1, with X_2 held at 1 in the first row and -1 in the second, and all other variables held at 0. The black solid line is the true value given by equation (20); the blue dotted line is the average fitted value over the set of Monte Carlo estimates. The blue interval represents the 2.5th and 97.5th percentiles of the sampling distribution of the fitted value over the set of Monte Carlo estimates.

With a binary outcome, there are more potential estimators to consider than in the previous case. The main question is whether to embed the basis regression within a logistic model. In this case, the model is set up as

$$\Pr(y_i = 1) = \frac{1}{1 + \exp(-\mu(x_i))} \equiv \nu(x_i), \tag{22}$$

and μ is estimated by running a logistic regression of $\mathbf{y}$ on a basis expansion of $\mathbf{X}$. However, notice that ν is continuous as long as μ is continuous, which suggests that we could directly estimate ν by running a least-squares regression of $\mathbf{y}$ on a basis expansion of $\mathbf{X}$. In theory, both approaches should recover the correct conditional expectation as the sample size and the number of terms in the basis expansion go to infinity. The logistic model may be more efficient when the true error distribution is logistic, such as in the present set of simulations, but there are few if any real-world cases where this distributional assumption is thought to be strictly true. The main reason to prefer a linear model is computational cost. For a dataset generated in this set of simulations, with $N = 1000$ and $\rho = 0$, fitting and bootstrapping 1000 times on recent hardware takes about 11 minutes for

a linear model with an adaptive LASSO penalty, compared to 75 minutes for a logistic model.[14] Computation times are even longer when a SCAD penalty is used: 231 minutes (3.9 hours) for a linear model and 696 minutes (11.6 hours) for a logistic model. In applications where the sample size is much larger, the additional time required to use a logistic model may be prohibitive. We also two consider two sets of initial weights for the adaptive LASSO penalty in logistic models: the logistic regression MLE and the OLS coefficient estimate. The advantage of the latter, particularly in small samples, is that separation (see Albert and Anderson 1984) cannot cause the estimate to diverge to infinity.

The results of this set of simulations are summarized in Table 3, which again shows the MISE and proportional reduction in error of each estimator. The performance of the oracle-penalized estimators relative to unpenalized basis regression and the naive model is essentially the same as in the previous set of simulations, so we turn our attention to other issues. The first question of interest is the relative performance of the linear and logistic basis regression models. As expected, the logistic basis regression models are more accurate than their linear counterparts, with about 5-10% better proportional reduction in error in each case. Since the simulation was set up to favor the logistic model, this probably represents an upper bound in the real-world performance improvement from its use. Average fitted values from the two types of model are plotted in Figure 2. The linear model recovers the overall shape of the relationship correctly, albeit with less precision than the logistic model. Estimates of $\mu(X)$ from the linear model do sometimes fall outside the unit interval, but only for relatively extreme values of X. On the whole, the results show that if an oracle-penalized logistic basis regression proves computationally infeasible, a researcher can still obtain reasonably accurate estimates of the conditional expectation from a linear version. The other issue we consider is which set of initial weights to use for the adaptive LASSO penalty with a logistic model. Perhaps surprisingly, we find that the MISE is typically lower when the OLS coefficients are used to compute the weights than when the logistic regression estimate is used, although the difference vanishes in large samples.

[14] Benchmarks were performed in R 2.15.2 on a machine with a 2.7 GHz Intel Core i7 processor.

N	ρ	ALASSO[a c]	ALASSO[b c]	ALASSO[b d]	SCAD[a]	SCAD[b]	MLE	naive	oracle	null
500	0.00	0.027 (0.67)	0.019 (0.77)	0.029 (0.64)	0.018 (0.78)	0.014 (0.84)	0.153 (−0.85)	0.025 (0.70)	0.005 (0.94)	0.082
500	0.25	0.026 (0.62)	0.019 (0.74)	0.029 (0.58)	0.018 (0.75)	0.014 (0.80)	0.147 (−1.08)	0.026 (0.63)	0.005 (0.93)	0.071
500	0.50	0.030 (0.44)	0.022 (0.59)	0.030 (0.44)	0.019 (0.65)	0.015 (0.71)	0.143 (−1.70)	0.024 (0.54)	0.005 (0.90)	0.053
500	0.75	0.028 (0.14)	0.021 (0.37)	0.024 (0.27)	0.019 (0.42)	0.016 (0.49)	0.136 (−3.16)	0.019 (0.41)	0.005 (0.84)	0.033
1000	0.00	0.014 (0.82)	0.008 (0.90)	0.013 (0.84)	0.010 (0.87)	0.005 (0.94)	0.048 (0.42)	0.023 (0.72)	0.002 (0.97)	0.082
1000	0.25	0.013 (0.81)	0.008 (0.88)	0.013 (0.81)	0.010 (0.86)	0.005 (0.93)	0.048 (0.32)	0.024 (0.66)	0.002 (0.97)	0.070
1000	0.50	0.013 (0.75)	0.009 (0.83)	0.014 (0.74)	0.010 (0.82)	0.006 (0.89)	0.048 (0.10)	0.023 (0.57)	0.003 (0.95)	0.053
1000	0.75	0.015 (0.55)	0.011 (0.67)	0.014 (0.57)	0.010 (0.69)	0.007 (0.78)	0.048 (−0.49)	0.017 (0.47)	0.002 (0.92)	0.032
5000	0.00	0.007 (0.91)	0.002 (0.98)	0.001 (0.98)	0.008 (0.91)	0.001 (0.99)	0.007 (0.92)	0.022 (0.74)	0.000 (0.99)	0.082
5000	0.25	0.006 (0.91)	0.002 (0.98)	0.001 (0.98)	0.007 (0.90)	0.001 (0.99)	0.007 (0.90)	0.023 (0.68)	0.000 (0.99)	0.070
5000	0.50	0.005 (0.90)	0.002 (0.97)	0.002 (0.97)	0.006 (0.89)	0.001 (0.98)	0.007 (0.87)	0.021 (0.60)	0.000 (0.99)	0.053
5000	0.75	0.005 (0.85)	0.002 (0.95)	0.002 (0.95)	0.005 (0.85)	0.001 (0.96)	0.007 (0.78)	0.016 (0.51)	0.000 (0.99)	0.032

Table 3. Results of the second Monte Carlo simulation, in the same format as Table 2.

[a] Linear model. [b] Logistic model. [c] Initial weights from OLS. [d] Initial weights from logistic regression MLE.

N	ρ	ALASSO[a]	SCAD[a]	OLS[a]	ALASSO[b]	SCAD[b]	OLS[b]	naive	oracle	null
200	0.00	0.123 (0.59)	0.112 (0.63)	0.316 (−0.06)	0.147 (0.51)	0.120 (0.60)	7.273 (−23.32)	0.129 (0.57)	0.046 (0.85)	0.299
200	0.25	0.113 (0.62)	0.109 (0.63)	0.301 (−0.01)	0.134 (0.55)	0.116 (0.61)	6.754 (−21.71)	0.118 (0.60)	0.045 (0.85)	0.297
200	0.50	0.109 (0.63)	0.108 (0.63)	0.326 (−0.12)	0.152 (0.48)	0.123 (0.58)	7.691 (−25.42)	0.111 (0.62)	0.046 (0.84)	0.291
200	0.75	0.097 (0.64)	0.090 (0.67)	0.305 (−0.12)	0.161 (0.41)	0.109 (0.60)	7.469 (−26.37)	0.095 (0.65)	0.046 (0.83)	0.273
500	0.00	0.066 (0.78)	0.073 (0.75)	0.099 (0.67)	0.095 (0.68)	0.085 (0.71)	0.420 (−0.41)	0.121 (0.59)	0.014 (0.95)	0.297
500	0.25	0.063 (0.79)	0.071 (0.76)	0.094 (0.68)	0.083 (0.72)	0.079 (0.73)	0.381 (−0.29)	0.110 (0.63)	0.014 (0.95)	0.296
500	0.50	0.062 (0.79)	0.067 (0.77)	0.095 (0.67)	0.081 (0.72)	0.074 (0.74)	0.425 (−0.47)	0.102 (0.65)	0.014 (0.95)	0.289
500	0.75	0.052 (0.81)	0.052 (0.81)	0.088 (0.67)	0.075 (0.72)	0.062 (0.77)	0.422 (−0.55)	0.087 (0.68)	0.015 (0.94)	0.271
1000	0.00	0.049 (0.83)	0.053 (0.82)	0.061 (0.79)	0.057 (0.81)	0.060 (0.80)	0.136 (0.54)	0.117 (0.60)	0.006 (0.98)	0.296
1000	0.25	0.048 (0.84)	0.053 (0.82)	0.058 (0.80)	0.055 (0.81)	0.058 (0.80)	0.124 (0.58)	0.107 (0.64)	0.006 (0.98)	0.295
1000	0.50	0.048 (0.83)	0.053 (0.82)	0.057 (0.80)	0.059 (0.80)	0.058 (0.80)	0.137 (0.53)	0.100 (0.65)	0.006 (0.98)	0.289
1000	0.75	0.041 (0.85)	0.042 (0.85)	0.052 (0.81)	0.053 (0.80)	0.045 (0.83)	0.133 (0.51)	0.085 (0.69)	0.006 (0.98)	0.271
5000	0.00	0.037 (0.88)	0.037 (0.88)	0.039 (0.87)	0.032 (0.89)	0.033 (0.89)	0.039 (0.87)	0.115 (0.61)	0.001 (1.00)	0.296
5000	0.25	0.036 (0.88)	0.036 (0.88)	0.036 (0.88)	0.030 (0.90)	0.033 (0.89)	0.035 (0.88)	0.105 (0.64)	0.001 (1.00)	0.295
5000	0.50	0.034 (0.88)	0.035 (0.88)	0.035 (0.88)	0.033 (0.89)	0.036 (0.88)	0.038 (0.87)	0.098 (0.66)	0.001 (1.00)	0.288
5000	0.75	0.030 (0.89)	0.032 (0.88)	0.031 (0.89)	0.031 (0.89)	0.034 (0.88)	0.037 (0.86)	0.083 (0.69)	0.001 (1.00)	0.270

Table 4. Results of the third Monte Carlo simulation, in the same format as Table 2.

[a] Power series basis, $d = 3$. [b] Power series basis, $d = 4$.

4.3 Non-Polynomial Equation

Our final set of Monte Carlo simulations examines a situation where the true conditional expectation function is not a polynomial. In this situation, basis regression can at best recover an approximation of μ, not the true model. We have two goals in mind with this set of trials. First, we want to examine the quality of the approximation that basis regression yields from a sample of typical size. Theoretical results guarantee that the expected estimate will converge to μ as the sample size and degree of the basis expansion increase, but it is important to have some idea of its reliability in finite samples. Second, we want to examine whether the advantages of the oracle-penalized models carry over to this setting. Since the oracle model is no longer a subset of the polynomial basis expansion, the features of the oracle property do not hold in this setting. However, penalization may nonetheless prove useful as a way to reduce instability in the estimates, a common problem for unregularized polynomial regression.

This set of simulations uses the following data-generating process, modeled after the one used by Hastie, Tibshirani and Friedman (2009, p. 327) in a similar test of the multivariate adaptive regression splines (MARS) estimator. There are now just five covariates, $X = (X_1, X_2, X_3, X_4, X_5)$, of which the first four are relevant and the last is noise. The model is given by

$$\begin{aligned}
\xi_1(X) &= \beta_{10} + \beta_{11}X_1 + \beta_{12}X_2 + \beta_{13}X_3 + \beta_{14}X_4 + \beta_{15}X_1X_2 + \beta_{16}X_3X_4, \\
\xi_2(X) &= \beta_{20} + \beta_{21}X_1 + \beta_{22}X_2 + \beta_{23}X_3 + \beta_{24}X_4 + \beta_{25}X_1X_3 + \beta_{26}X_2X_4, \\
\mu(X) &= \alpha_0 + \alpha_1\Phi(\xi_1(X)) + \alpha_2\Phi(\xi_2(X)), \\
y_i &= \mu(x_i) + \epsilon_i,
\end{aligned} \tag{23}$$

where $\Phi(\cdot)$ is the CDF of a standard normal distribution.[15] The error ϵ_i is normally distributed with mean zero and variance 0.3. As before, the distribution of X is joint normal; each variable has unit variance and is correlated with all others at the level ρ. We examine the same values of ρ and the sample size N as in Section 4.1. A rough interpretation of equation (23) is that $\xi_1(X)$ and $\xi_2(X)$ represent latent traits of an individual, and $\mu(X)$ is the difference in the individual's percentiles in the two traits.

[15] The coefficients are fixed at $\alpha \equiv (\alpha_0, \alpha_1, \alpha_2) = (0, 1, -1)$, $\beta_1 \equiv (\beta_{10}, \ldots, \beta_{16}) = (1, 1, 1, 1, 1, -1)$, and $\beta_2 \equiv (\beta_{20}, \ldots, \beta_{26}) = (1, -1, 1, -1, -1, 1)$.

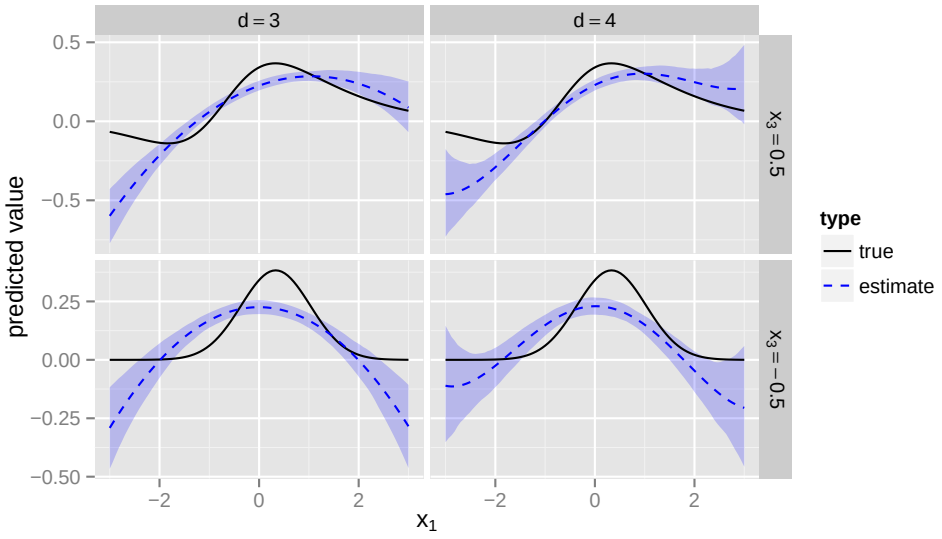

Figure 3. Illustration of fitted values in the third Monte Carlo simulation, in the same format as Figure 2. All values are taken from adaptive LASSO models in the set of trials with $N = 5000$ and $\rho = 0$. The first column represents a fit using a power series basis expansion of degree 3; the second uses an expansion of degree 4. X_3 is held at 0.5 in the first row and -0.5 in the second, while the other variables are held at $X_2 = 0.5$, $X_4 = 0$, and $X_5 = 0$ throughout.

The set of estimators that we employ is mainly the same as in Section 4.1, with two differences. First, for the basis regression models, we use power series basis expansions of both degree $d = 3$ and $d = 4$. Because there are five covariates, there are 56 basis terms included when $d = 3$ and 126 included when $d = 4$. Second, because equation (23) is not a typical linear or logistic regression equation, it is not quite as straightforward to estimate the oracle model as in the previous rounds of simulations. We use nonlinear least squares to estimate α, β_1, and β_2 from equation (23), implemented with the `nls()` function in R.

The results of this set of simulations are summarized in Table 4. Once again, we see that the oracle-penalized basis regression estimators outperform a naive linear model in almost all cases, with the exception of the degree-4 models with $N = 200$. Since the true conditional expectation is a highly interactive nonlinear function, it is not surprising that a simple linear model yields a poor approximation. On the other hand, the differences between the penalized estimators and the ordinary basis regression are less pronounced than in the previous simulations. Although the MISE of the penalized estimators is always lower than that of ordinary basis regression, the difference nearly vanishes in large samples. This suggests that when enough data is available, approximation error is a greater concern for basis regression than is high variance. Nonetheless, seeing as the

penalized estimators still have slightly lower error in large samples (and significantly less in small samples), their only drawback appears to be greater computational cost.

As expected, the relative performance of basis expansions of degree $d = 3$ and $d = 4$ depends on sample size: the larger expansion is preferred only when N is large. In particular, the MISE of the degree-3 estimators is substantially lower than that of the degree-4 models for $N \leq 500$, roughly equal (but still slightly lower) for $N = 1000$, and slightly higher for $N = 5000$. The differences are much more pronounced for unpenalized basis regression than in the adaptive LASSO and SCAD models. These results highlight the importance of using cross-validation to select the degree of the basis expansion in particular applications; recall that the cross-validation error in equation (15) is an estimate of the MISE. For an illustration of fitted values from the adaptive LASSO under both $d = 3$ and $d = 4$, see Figure 3. Near the center of the distribution of the covariates, the two estimators yield essentially the same estimate of $\mu(X)$ on average. However, for more extreme values of X, the higher-degree estimate better matches the true conditional expectation function. Neither approximation is perfect, as expected, but both retrieve the main qualitative features of μ in the particular covariate profiles illustrated.

5 Reassessing the Capitalist Peace

In this section, we illustrate our oracle-penalized basis regression technique by using it to reanalyze a prominent empirical finding in political science: Gartzke's (2007) result that the democratic peace is a spurious consequence of capitalist economic factors. Gartzke's main result is that the estimated effect of joint democracy on dispute occurrence is statistically insignificant (though still negative) when measures of economic interdependence and development are included in the regression specification. This comes out of a set of logistic regression models with restrictive linear specifications, in which almost all of the variables are assumed to have an unconditionally monotonic relationship with the probability of a dispute occurring. Because of the large sample size, which ranges from about 165,000 to 282,000 depending on the exact specification, these models are an ideal candidate for replication with nonparametric methods. If the true form of the relationship between the covariates and the probability of a dispute is monotonic and non-interactive, there is enough information in the data that even a very flexible model is likely to capture it. By the same

token, if the estimated model suggests that the relationship cannot be captured in a restrictive linear specification, this result cannot be easily dismissed as overfitting or high variability.

Our replication uses the variables from Gartzke's Model 2 (p. 177), the simplest specification in which the "capitalist peace" result arises. We will briefly review the model here; for full information on coding and data sources, see Gartzke (2007, pp. 174–176) The dataset contains 174,548 observations on dyad-years between 1950 and 1992, and the dependent variable is the occurrence of a militarized international dispute. The covariates include two "capitalist" variables for each dyad. The first is openness, the lower of the two states' scores on an International Monetary Fund index of trade openness. The second is dependence, the lower of the bilateral trade-to-GDP ratios between the two members of the dyad. The "democratic" variables are the Polity scores for the two states in the dyad, separated into lower and higher score. Other covariates included are an indicator for whether the states are contiguous, the distance between them, an indicator for whether either is a major power, an indicator for whether they are in an alliance, and the "capability ratio" of the stronger state's Correlates of War capability index to the weaker state's. The model also includes indicators for when both states in the dyad are in the same of seven regional categories: Asia, Europe, North Africa/Middle East, sub-Saharan Africa, North America, South America, and the West Pacific.[16] Finally, the model deals with temporal dependence via Beck, Katz and Tucker's (1998) method of including a spline transformation of years since the last conflict. When the model is rerun using Carter and Signorino (2010) method of including a polynomial transformation of peace years, the results are substantively identical. Moreover, the non-nested model comparison tests of both Vuong (1989) and Clarke (2006) find that the model with polynomial peace years fits significantly better than the one with spline peace years. We thus use the model with polynomial peace years as the baseline for comparison to our nonparametric estimates.

We replicate the analysis using regression with an adaptive LASSO penalty on a power series basis. Although the response variable (conflict occurrence) is binary, we use a linear model, as the large sample size makes computation of a penalized logistic model prohibitive.[17] Two issues that

[16] Gartzke drops the West Pacific indicator from the analysis "to avoid a dummy variable trap" (p. 176, fn. 46). However, because there are observations that do not fall into any of the regional categories—the dyads that contain countries in different regions—this step is unnecessary. In fact, it amounts to assuming that the intercept is the same for West Pacific dyads as for mixed-region dyads. When we rerun the original analysis including the West Pacific indicator, all results are virtually identical.

[17] For a discussion of linear versus logistic models for binary response variables, see Section 4.3.

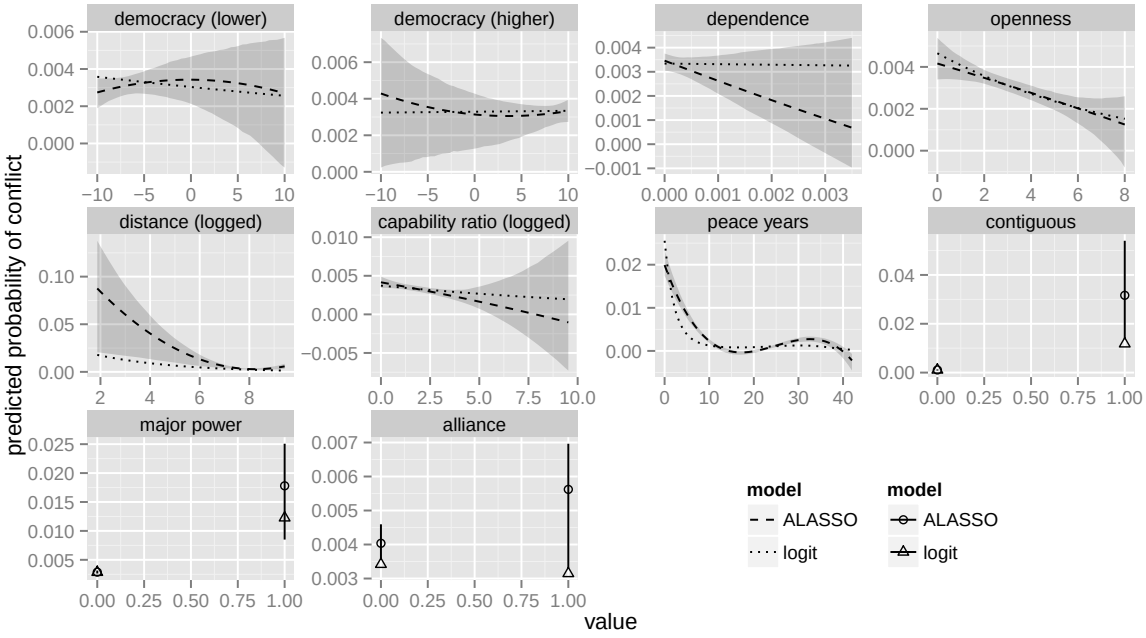

Figure 4. Average relationships between covariates and the predicted probability of a militarized international dispute, calculated using the observed-value method. The intervals depicted are the 95% bootstrap confidence intervals for the estimates from the adaptive LASSO–penalized basis regression model.

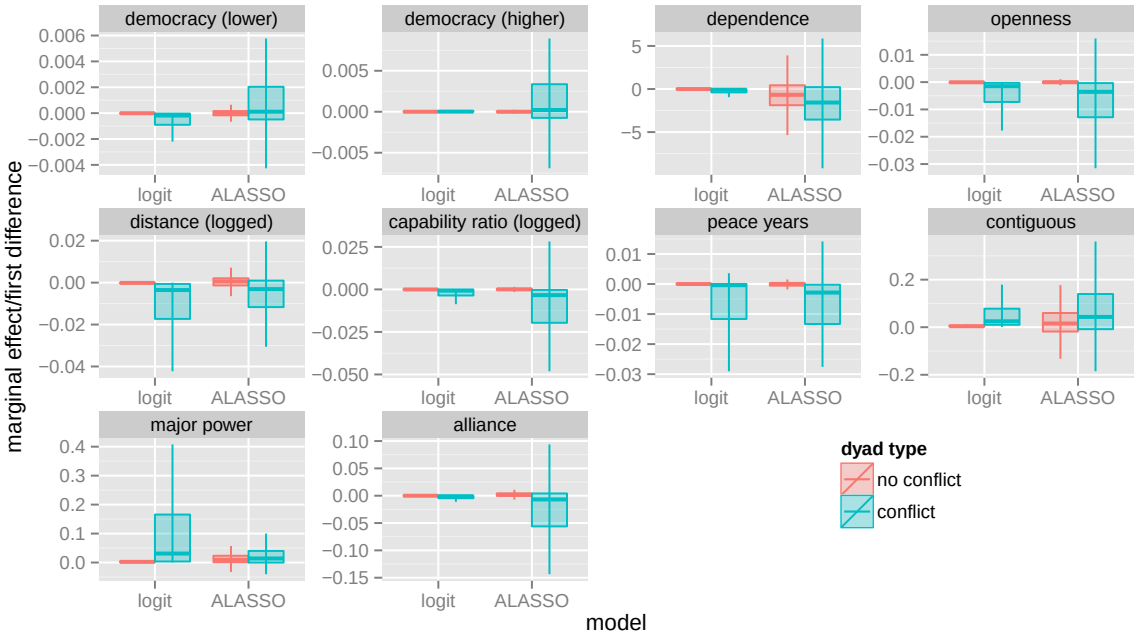

Figure 5. Box plots of the distribution of estimated effects of each covariate on the probability of a militarized international dispute. The values represent marginal effects for the first seven (continuous) variables, and first differences for the last three (binary) variables. To preserve legibility, outliers are not plotted.

arise are how to treat the regional and duration controls. To avoid adding hundreds of terms to the model, we simply include the regional indicators linearly; this is equivalent to assuming (as in the original model) that regional membership only affects the baseline level of conflict. For duration dependence, we consider two alternatives: including peace years in the full basis expansion, and including only a cubic polynomial in peace years without allowing it to be interacted with the substantive covariates. We use 10-fold cross-validation to choose between these two setups, as well as to select the tuning parameters. The model that minimizes the cross-validation error (see equation (15)) uses a power series basis of degree $d = 3$ and a penalty factor of $\lambda = 0.019$, and it includes peace years in the full basis expansion. We therefore fit this model on the full dataset and then run 1,000 iterations of the nonparametric bootstrap to estimate the standard errors.[18]

To summarize the results of the estimated basis regression model, we begin with a familiar approach: plotting the average relationship between each covariate and the predicted probability of conflict. Political scientists often construct such plots by constructing an "average observation" where each covariate is held at its mean or median; the average relationship is then calculated by inserting each value of a given covariate (or, for continuous variables, a grid of values) into this observation. However, as Hanmer and Kalkan (2012) persuasively argue, this approach is flawed. In a nonlinear model, such as our nonparametric model or even ordinary logistic regression, the relationship between x and y within the "average observation" is not equivalent to the average relationship across the population. Therefore, to plot these relationships, we use the observed-value approach: "holding each of the other independent variables at the observed values for each case in the sample, calculating the relevant predicted probabilities or marginal effect for each case, and then averaging over all of the cases" (Hanmer and Kalkan 2012, p. 264). Confidence intervals around the relationship can then be calculated by repeating this procedure with each set of bootstrap estimates and computing the relevant order statistics.

These plots of average relationships are useful as a first-order summary of the model results, but they must be interpreted carefully. Most importantly, the effect that a change in one variable would have on the expected outcome in a particular case may differ substantially from the population average. For example, a variable could appear to have a slight negative relationship with

[18] For simplicity, we use simple random resampling. To allow for groupwise dependence in errors, a block bootstrap could be used instead.

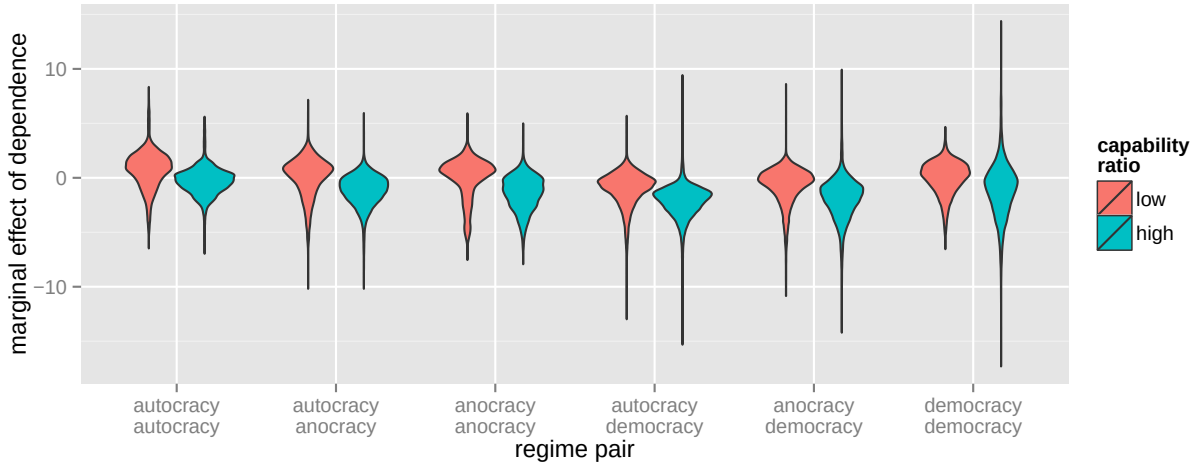

Figure 6. Violin plot of the relationship between regime type, capability ratio, and the distribution of the estimated marginal effect of dependence on the probability of conflict.

the outcome on average when a marginal increase would have a strong negative effect in 50% of observations, no effect in 25%, and a strong positive effect in the remaining 25%. It is important to distinguish that kind of variable from one that has a slight negative effect in all cases, even though the average relationship will look the same for both. To this end, we advocate plotting the *distribution* of marginal effects. This entails computing $\partial\hat{\mu}(x_i)/\partial X_m$ via equation (12) for each observation i and covariate X_m. For binary variables, a more appropriate measure is the difference in the predicted probability between when the variable is set to 1 and when it is set to 0.

The distributions of estimated marginal effects and first differences from both the original model and the nonparametric replication are plotted in Figure 5. Observations are divided into two groups depending on whether the dyad ever experiences conflict during the time period studied, putting 6,442 dyad-years in the "conflict" category and 168,106 in the "no conflict" category. The first result of note is that, for both models, the range of estimated effects is wider in the subsample of conflict-prone dyads. This is essentially true by construction for the logit model, whose functional form implies that the marginal effects of all variables increase as the predicted probability approaches 0.5. The more substantively interesting finding is that, according to our nonparametric model, many of the key variables have a conditionally monotonic relationship with the probability of conflict. In other words, these variables have a positive effect on the likelihood of a dispute in many cases, and a negative effect in many others. Within the subsample of conflict-prone dyads, the interquartile

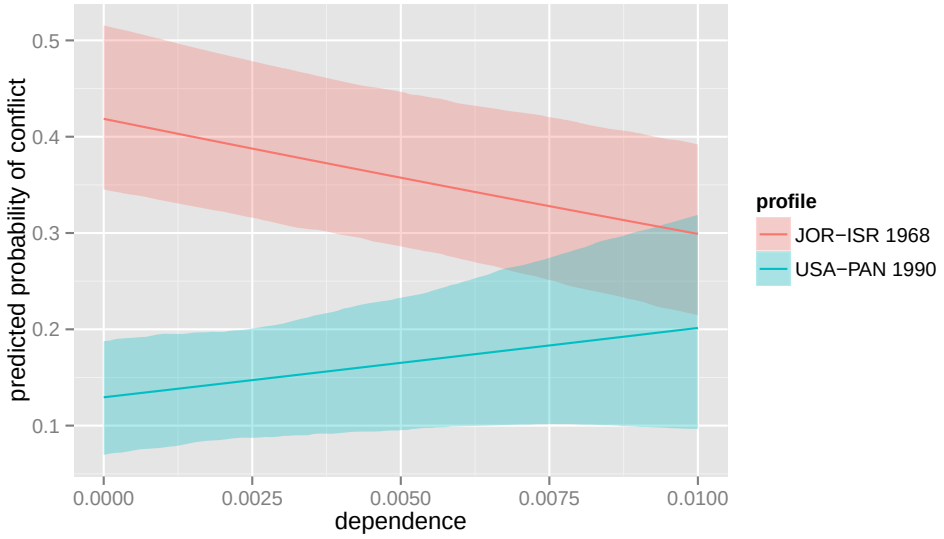

Figure 7. Estimated relationship between dependence and the probability of conflict when all other variables are set to the values in the Jordan–Israel 1968 and U.S.A.–Panama 1990 observations. The intervals depicted are 95% bootstrap confidence intervals.

ranges of the marginal effects of dependence and the democracy measures contain 0, meaning each has a positive effect on conflict in at least a quarter of cases. Among the key variables in the analysis, only openness has a negative effect on the probability of a dispute in more than 75% of conflict-prone dyad-years. These results of course differ from the original logit model, in which it is assumed *a priori* that the sign of each key variable's effect is the same in all cases.

On further inspection of our results, it becomes clear that the effects of "democratic peace" and "capitalist peace" variables cannot be neatly disentangled from one another. On the contrary, the effect of trade dependence on the likelihood of conflict is in part a function of the regime types represented in the dyad. To illustrate this, we plot the distribution of the estimated marginal effect on dependence across different regime-type pairings in Figure 6. Within each regime pair, the distribution is plotted separately for dyads with logged capability ratios below and above the sample median (approximately 1.6). The most apparent result is that greater dependence usually decreases the risk of conflict in dyads with a large imbalance in capabilities—which, as shown in Figure 4, are less prone to conflict in the first place—but typically increases the chance of conflict in more balanced dyads. However, these effects vary with regime type. Most notably, greater trade dependence more often has a pacifying effect in dyads including exactly one democracy than in other cases. In addition, among dyads with a high capability ratio, dependence is relatively likely

to increase the chance of conflict in fully autocratic and fully democratic dyads. As a further illustration of variation in the estimated effect of dependence, Figure 7 plots the estimated relationship between trade dependence and the chance of conflict for two dyad-years in the dataset: Jordan–Israel 1968 and U.S.A.–Panama 1990. Both are one year removed from a dispute (hence the high baseline probabilities of conflict), but the former is autocratic-democratic with fairly even capabilities, while the latter is jointly democratic with mismatched capabilities. According to the estimated model, additional trade dependence would have decreased the risk of conflict between Jordan and Israel in 1968, whereas it would have slightly increased the chance of a dispute between the United States and Panama in 1990.

In summary, our findings tell a more nuanced story about the relationship between regime type, economic factors, and peace than the original model does. Although our model suggests that both interdependence and joint democracy reduce conflict on average, it also implies that many jointly democratic dyads could decrease their chance of conflict by reducing their trade dependence. This element of the capitalist peace does not appear to operate within the countries covered by the democratic peace. More broadly, this application illustrates the importance of flexible functional form modeling, especially for estimation with large datasets like this one. Our results do not just differ from the original model; they simply could not have been found under its restrictive specification. The first-order, average effect—which is all that a typical linear regression specification can capture—may mask a great deal of variation in the substantive relationships of interest. Flexible, nonparametric modeling can help prevent researchers from drawing unconditional or global conclusions that are not borne out in the data. Moreover, the discovery of exceptions to the average relationship can provide fruitful avenues for future inquiry.

6 Concluding Remarks

We have introduced a new method for flexible form estimation and demonstrated its utility in both simulations and an empirical application. More broadly, we hope to spur the wider adoption of nonparametric modeling by political scientists. Although useful as a first approximation, standard modeling techniques cannot capture the rich, complex relationships that prevail in the social world. Nonparametric methods can make existing analyses more accurate by preventing against functional

form misspecification bias, and they can help spur further research by highlighting nonlinear relationships that a researcher may not have expected. Moreover, as the amount of data available to practitioners continues to increase, so do the potential applications of data-intensive methods like nonparametric regression. In companion papers, we argue that nonparametric methods can be used to improve studies of sample-selection bias (Kenkel and Signorino 2012*a*) and the empirical analysis of predictions from formal models (Kenkel and Signorino 2012*b*).

To facilitate the application of our new method, we have developed the R package `polywog` (Kenkel and Signorino 2013), which implements most of the techniques described in this paper. The package contains a function for fitting polynomial basis regression models with an adaptive LASSO or SCAD penalty. Users may select the degree of the basis expansion or have it automatically selected via K-fold cross-validation. The bootstrap is implemented with optional parallel processing, so that users with multicore machines or clusters can save on computation time. The package also provides functions for the analysis of fitted models, including plots of average relationships and of the distribution of estimated marginal effects. The package is under active development and is available via CRAN at `http://cran.r-project.org/web/packages/polywog/index.html`.

References

Albert, A and J A Anderson. 1984. "On the Existence of Maximum Likelihood Estimates in Logistic Regression Models." *Biometrika* 71(1):1–10.

Andrews, Donald W K. 1991. "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Models." *Econometrica* 59(2):307–345.

Beck, Nathaniel, G King and L Zeng. 2000. "Improving Quantitative Studies of International Conflict: A Conjecture." *The American Political Science Review* 94(1):21–35.

Beck, Nathaniel, Jonathan N Katz and Richard Tucker. 1998. "Taking Time Seriously: Time-Series-Cross-Section Analysis with a Binary Dependent Variable." *American Journal of Political Science* 42(4):1260–1288.

Beck, Nathaniel and Simon Jackman. 1998. "Beyond Linearity by Default: Generalized Additive Models." *American Journal of Political Science* 42(2):596–627.

Breheny, Patrick and Jian Huang. 2011. "Coordinate Descent Algorithms for Nonconvex Penalized Regression, with Applications to Biological Feature Selection." *The Annals of Applied Statistics* 5(1):232–253.

Breiman, Leo. 1996. "Bagging predictors." *Machine Learning* 140:123–140.

Carter, David B and Curtis S Signorino. 2010. "Back to the Future: Modeling Time Dependence in Binary Data." *Political Analysis* 18(3):271–292.

Chatterjee, A and S N Lahiri. 2011. "Bootstrapping Lasso Estimators." *Journal of the American Statistical Association* .

Clarke, Kevin A. 2006. "A Simple Distribution-Free Test for Nonnested Model Selection." *Political Analysis* 15(3):347–363.

Derksen, Shelley and H J Keselman. 1992. "Backward, Forward, and Stepwise Automated Subset Selection Algorithms: Frequency of Obtaining Authentic and Noise Variables." *British Journal of Mathematical and Statistical Psychology* 45:265–282.

Efromovich, Sam. 1999. *Nonparametric Curve Estimation: Methods, Theory, and Applications.* Springer-Verlag.

Efron, Bradley and Robert Tibshirani. 1993. *An Introduction to the Bootstrap.* Chapman & Hall/CRC.

Eubank, Randall L. 1999. *Nonparametric Regression and Spline Smoothing.* 2 ed. CRC Press.

Fan, Jianqing and Runze Li. 2001. "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties." *Journal of the American Statistical Association* 96(456):1348–1360.

Friedman, Jerome H, Trevor Hastie and Robert Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33(1).

Gallant, Ronald A. 1981. "On the Bias in Flexible Functional Forms and an Essentially Unbiased Form: The Fourier Flexible Form." *Journal of Econometrics* 15:211–245.

Gartzke, Erik. 2007. "The Capitalist Peace." *American Journal of Political Science* 51(1):166–191.

Hainmueller, Jens and Chad Hazlett. 2012. "Kernel Regularized Least Squares: Moving Beyond Linearity and Additivity Without Sacrificing Interpretability." Typescript, Massachusetts Institute of Technology.

Hanmer, Michael J and Kerem Ozan Kalkan. 2012. "Behind the Curve: Clarifying the Best Approach to Calculating Predicted Probabilities and Marginal Effects From Limited Dependent Variable Models." *American Journal of Political Science* 57(1):263–277.

Hastie, Trevor and Robert Tibshirani. 1990. *Generalized Additive Models.* Chapman & Hall/CRC.

Hastie, Trevor, Robert Tibshirani and Jerome H Friedman. 2009. *The Elements of Statistical Learning.* 2 ed. Springer.

Hurvich, Clifford M and Chih-Ling Tsai. 1990. "The Impact of Model Selection on Inference in Linear Regression." *The American Statistician* 44(3):214–217.

Keele, Luke John. 2008. *Semiparametric Regression for the Social Sciences.* Wiley.

Kenkel, Brenton and Curtis S Signorino. 2012*a*. "An Alternative Solution to the Heckman Selection Problem: Selection Bias as Functional Form Misspecification." Typescript, University of Rochester.

Kenkel, Brenton and Curtis S Signorino. 2012*b*. "Data Mining for Theorists." Typescript, University of Rochester.

Kenkel, Brenton and Curtis S Signorino. 2013. *polywog: Bootstrapped Basis Regression with Oracle Model Selection*. 0.3-0 ed.

King, Gary. 1986. "How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science." *American Journal of Political Science* 30(3):666–687.

King, Gary, Michael Tomz and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44(2):347–361.

Kmenta, Jan. 1986. *Elements of Econometrics*. 2 ed. Macmillan.

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

Newey, Whitney K. 1994. "Series Estimation of Regression Functionals." *Econometric Theory* 10(1):1–28.

Pagan, Adrian and Aman Ullah. 1999. *Nonparametric Econometrics*. Cambridge University Press.

Signorino, Curtis S and Kuzey Yilmaz. 2003. "Strategic Misspecification in Regression Models." *American Journal of Political Science* 47(3):551–566.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society, Series B* 58(1):267–288.

Vuong, Quang H. 1989. "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses." *Econometrica* 57(2):307–333.

Wand, Jonathan. 2011. "More than a Science of Averages: Testing Theories Based on the Shapes of Relationships." Typescript, Stanford University,.

Wand, Jonathan. 2012. "Testing Competing Theories with Shape Constrained Inference." Typescript, Stanford University,.

Zou, Hui. 2006. "The Adaptive Lasso and Its Oracle Properties." *Journal of the American Statistical Association* 101(476):1418–1429.