Online Appendix:
Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach

Jens Hainmueller – Massachusetts Institute of Technology
Chad Hazlett – Massachusetts Institute of Technology

August 2013

Abstract

This online appendix presents various additional results, explanations, proofs, and simulations referenced in the main text.

Jens Hainmueller, Department of Political Science, 77 Massachusetts Avenue, Cambridge, MA 02139. E-mail: jhainm@mit.edu. Jens Hainmueller, Department of Political Science, 77 Massachusetts Avenue, Cambridge, MA 02139. E-mail: jhainm@mit.edu. Chad Hazlett, Department of Political Science, 77 Massachusetts Avenue, Cambridge, MA 02139. E-mail: hazlett@mit.EDU.

1. Illustration of the Gaussian Superposition View

The text describes three primary views that are equally valid interpretations of KRLS: a ridge regression in an infinite-dimensional transform of the original features, the similarity-based view, and the Gaussian superposition view. Here we provide additional illustrations to help develop the intuition of the Gaussian superposition view, which we could not accommodate in the main text.

The Gaussian Superposition view interprets KRLS as a process of building up the solution surface through placing Gaussian curves or mounds over each observation in the dataset, then scaling them such that the summated surface of these overlapping mounds approximates the response data. To build this intuition, consider a simple dataset in which $x = [1, 2, 3, 4]$, and $y = [-2, 0, 1, 1]$. We begin by choosing the x_i values in the observed data, and placing Gaussians over them. The unscaled Gaussians are visualized in the left panel of Figure A.1. Next, we use the KRLS algorithm to chose the scaling coefficients c_i. The right panel in Figure A.1 shows each Gaussian after re-scaling by the corresponding c_i, as dotted lines. It also shows the final step of the process, the summated surface, as a solid line. The same logic generalizes to multidimensional functions. Note that the fitted function passes through the original data points, shown as dots, though this is not guaranteed in noisier data. In addition, the continuity of the Gaussians allows us to estimate the value of $f(x)$ for any other x in-between the existing observations. This representation also makes it easy to see why we can compute derivatives at each point: since y is formed by a sum of differentiable functions, it is itself differentiable.

Regularization plays its essential role through the choice of the scaling coefficients. Instead of choosing relatively small values of c_i as KRLS did in this example, one could have instead chosen very large but opposing values of c_i, corresponding to very large rescaled Gaussians that mostly offset each other. At the observations, such functions can fit the data similarly well or better than the "small c_i" solution, but at the expense of (undesirable) "wiggliness" at locations in-between the observed points. Through regularization, KRLS avoids choosing these large offsetting coefficients by explicitly punishing models with larger (squared) coefficients. The norm $c^T c$ would be a reasonable choice for establishing this penalty (and works well in practice), corresponding to the notion of a ridge regression in the columns of K. KRLS instead uses the norm $c^T K c$, corresponding to a ridge regression instead in the feature space associated with the kernel. More intuitively, using $c^T K c$ instead of $c^T c$ has the effect of more heavily punishing large coefficients when they correspond to observations closer to each other in X (i.e. with more heavily overlapping Gaussians).

2. Regularization as a Bayesian Prior

In the main text we motivate Tikhonov regularization as a natural strategy for achieving stability and generalizability outside the trained sample. However, regularization is also justifiable as the maximum a posteriori estimator after imposing a prior that models with greater complexity are expected to be less likely. Let $y_i = f(x_i) + \epsilon_i$, with $\epsilon \sim N(0, \sigma_\epsilon^2)$. Let our prior be that low-complexity functions are more likely than high complexity ones, and specifically that the prior probability of observing a model with complexity $||f||_K^2$ is proportional to $e^{-\alpha||f||_K^2}$ where α is some positive constant and $||f||_K^2$ is the L_2 norm in the feature space associated with K, i.e. $c^T Kc$. Then, the posterior probability of observing the given dataset is proportional to:

$$\prod_{i=1}^{N} e^{-\frac{(y_i - f(x_i))^2}{\sigma_\epsilon^2}} \; e^{-\alpha||f||_K^2} \tag{1}$$

The maximum of this posterior can be found by minimizing its negative log:

$$\underset{f \in H}{\operatorname{argmin}} \sum_i \frac{(y_i - f(x_i))^2}{\sigma_\epsilon^2} + \alpha||f||_K^2 \tag{2}$$

or equivalently by minimizing $\sum_i (f(x_i) - y_i)^2 + \lambda||f||_K^2$ with $\lambda = \sigma_\epsilon^2 \alpha$, which is simply Tikhonov regularization.

3. Solving for $c^\star$

We wish to choose $c^\star$ to satisfy:

$$\underset{c \in \mathbb{R}^D}{\operatorname{argmin}} \, (y - Kc)^T (y - Kc) + \lambda c^T Kc \tag{3}$$

This is easily achieved by solving the first order conditions:

$$\begin{aligned}
J &= (y - Kc)^T (y - Kc) + \lambda c^T Kc \\
\frac{\partial J}{\partial c} &= -2K(y - Kc) + 2\lambda Kc \\
0 &= -2K(y - Kc) + 2\lambda Kc \\
y &= c(K + \lambda I) \\
c^\star &= (K + \lambda I)^{-1} y
\end{aligned} \tag{4}$$

4. Choice of Kernel Bandwidth

The key result that motivates our approach of choosing the kernel bandwidth automatically is that using the standardized data, the average distance between observations is simply two times the number of dimensions. This is shown as follows: let $x_i^{(a)}$ designate the i^{th} observation of the a^{th} independent variable then the expected squared difference between any two pairs of observations is

$$\begin{aligned}
S^2 &= E[||x_j - x_i||^2] \\
&= E[(x_j^{(a)} - x_i^{(a)})^2 + (x_j^{(b)} - x_i^{(b)})^2 + \ldots + (x_j^{(D)} - x_i^{(D)})^2] \\
&= E[(x_j^{(a)})^2 + (x_i^{(a)})^2 + 2(x_j^{(a)})(x_i^{(a)})] + \ldots + E[(x_j^{(D)})^2 + (x_i^{(D)})^2 + 2(x_j^{(D)})(x_i^{(D)})] \\
&= \sigma_a^2 + \sigma_a^2 + 2cov(x_j^{(a)}, x_i^{(a)}) + \ldots + \sigma_D^2 + \sigma_D^2 + 2cov(x_j^{(D)}, x_i^{(D)}) \\
&= 2\sum_d \sigma^2 = 2D \qquad (5)
\end{aligned}$$

where we use the fact that, after standardization, each variable has mean zero and equal variance $\sigma_a^2 = \sigma_b^2 = ... = \sigma_D^2$, and also that the data are i.i.d. such that the pairwise covariances are zero. This does not depend on the correlation of the different variables within X, only on the i.i.d. nature of each draw of x_i across the observations.

Now, recall that the Gaussian kernel is $k(x_j, x_i) = e^{-\frac{||x_j - x_i||^2}{\sigma^2}}$. Since the average squared distance between two points on the standardized data is simply two times the number of dimensions, the average value for the numerator of the exponent, $||x_j - x_i||^2$, is $2D$. Choosing σ^2 to be proportional to D ensures a reasonable scaling of these distances on average, regardless of D. What exactly the constant of proportionality should be is an open question, though $\sigma^2 = 1D$ seems to be a reliable choice in practice. In the many test cases that we have examined, the resulting distribution of K values from this choice has been reasonable, some observations to be considered similar (values near 1), some to be very dissimilar (near 0), and a spread out distribution in between.

5. Proof of Theorems 1-2: Unbiasedness

First, we prove Theorem 1, the unbiasedness of the choice coefficients for the target choice coefficients in the available space of functions.

Assumption 1 (Functional Form) *The target function we seek to estimate falls in the space of functions representable as $y^\star = Kc^\star$ and we observe a noisy version of this, $y_{obs} = y + \epsilon$.*

Assumption 2 (Zero Conditional Mean) $E[\epsilon|X] = 0$, *which implies that* $E[\epsilon|K_i] = 0$ *(where K_i designates the i^{th} column of K) since K is a deterministic function of X.*

Theorem 1 (Unbiasedness of choice coefficients) *Under assumptions 1–2,* $E[\hat{c}^\star|X] = c^\star$.

$$\begin{aligned}
E[\hat{c}^\star|X] &= E[(K+\lambda I)^{-1} y_{obs}] \\
&= E[(K+\lambda I)^{-1}(y+\epsilon)] \\
&= E[(K+\lambda I)^{-1} y] + E[(K+\lambda I)^{-1}\epsilon] \\
&= E[(K+\lambda I)^{-1} y] + (K+\lambda I)^{-1} E[\epsilon|K] \\
&= E[(K+\lambda I)^{-1} y] \\
&= c^\star
\end{aligned} \tag{6}$$

The unbiasedness of the fitted $\hat{y}$ vector follows immediately:

Theorem 2 (Unbiasedness of fitted values) *Under assumptions 1–2,* $E[\hat{y}] = y^\star$.

$$E[\hat{y}|X] = E[K\hat{c}^\star] = KE[\hat{c}^\star] = Kc^\star = y^\star \tag{7}$$

We note that the estimates of $\hat{c^\star}$ are unbiased for $c^\star$, not for c. In other words, we take as part of the correct specification requirement (assumption 1) that the target function we seek to estimate unbiasedly is one which has taken the complexity punishment into account. This is consistent with the view that we utilize regularization precisely because we think regularized functions are more likely to represent the generalizable, stable, or useful relationship between X and y. Put differently, the unbiasedness result obtains only conditional on a chosen level of regularization.

6. Proof of Lemma 1

Lemma 1 gives a closed form expression for the variance-covariance matrix of $\hat{c}$.

Assumption 3 (Spherical Errors) *The errors are homoscedastic and have zero serial correlation, such that* $E[\epsilon\epsilon^T|X] = \sigma_\epsilon^2 I$.

Lemma 1 (Variance of choice coefficients) *Under assumptions 1,2, and 3, the variance of the choice coefficients is given by* $\text{Var}[\hat{c}^\star|X,\lambda] = \sigma_\epsilon^2(K+\lambda I)^{-2}$.

$$\begin{aligned}
\text{Var}[\hat{c}^\star|X,\lambda] &= E[(\hat{c}^\star - E(\hat{c}^\star))^2|X,\lambda] \\
&= E[(K+\lambda I)^{-1}\epsilon\epsilon^T(K+\lambda I)^{-1}|K] \\
&= (K+\lambda I)^{-1}E[\epsilon\epsilon^T|K](K+\lambda I)^{-1} \\
&= (K+\lambda I)^{-1}\sigma_\epsilon^2 I(K+\lambda I)^{-1} \\
&= \sigma_\epsilon^2(K+\lambda I)^{-2} \qquad (8)
\end{aligned}$$

We note that assumption 3 is made for convenience here, in parallel to the spherical error assumption often made under linear models. More complicated forms for the covariance matrix of the errors can be introduced for $E[\epsilon\epsilon^T|X]$, which would allow computation of other standard error estimators. A particularly useful extension will be the development of cluster-robust standard errors through an appropriate choice of $E[\epsilon\epsilon^T|X]$.

7. Proof of Theorem 3: Consistency

Assumption 4 (Regularity Condition I) *Let (i)* $\lambda > 0$ *and (ii) as* $N \to \infty$, *for eigenvalues of* K *given by* a_i, $\sum_i \frac{a_i}{a_i+\lambda}$ *grows slower than* N *once* $N > M$ *for some* $M < \infty$.

Here we establish Theorem 3, the consistency of the KRLS estimator.

Theorem 3 (Consistency) *Under assumptions 1-4,* $E[\hat{y}_i|X] = y_i^\star$ *and* $\underset{N\to\infty}{\text{plim}}\ \text{Var}[\hat{y}|X,\lambda] = 0$ *and the estimator is therefore consistent with* $\underset{N\to\infty}{\text{plim}}\ \hat{y}_{i,N} = y_i^\star$ *for all* i.

The proof is as follows. First, since K is symmetric and positive semi-definite, we can decompose it into its eigenvectors, V, and eigenvalues on the diagonal of A. We can now rewrite the variance-covariance matrix of $\hat{y}^\star$ as:

$$\begin{aligned}
\mathrm{Var}[\hat{y}|X,\lambda] &= K\ \mathrm{Var}[c^\star]\,K^T \\
&= K[\sigma_\epsilon^2 I(K+\lambda I)^{-2}]K^T \\
&= \sigma_\epsilon^2 KK(K+\lambda I)^{-1}(K+\lambda I)^{-1} \\
&= \sigma_\epsilon^2 (K(K+\lambda I)^{-1})^2 \\
&= \sigma_\epsilon^2 (VAV^TV(A+\lambda I)^{-1}V^T)^2 \\
&= \sigma_\epsilon^2 V(A(A+\lambda I)^{-1})^2 V^T \\
&= \sigma_\epsilon^2 V \begin{bmatrix} (\frac{a_1}{a_1+\lambda})^2 & 0 & \dots & 0 \\ 0 & (\frac{a_2}{a_2+\lambda})^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & (\frac{a_N}{a_N+\lambda})^2 \end{bmatrix} V^T \qquad (9)
\end{aligned}$$

Where a_i designates the i^{th} diagonal element of A (i.e. the i^{th} eigenvalue of K). For convenience, define matrix M as follows:

$$M = \begin{bmatrix} (\frac{a_1}{a_1+\lambda}) & 0 & \dots & 0 \\ 0 & (\frac{a_2}{a_2+\lambda}) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & (\frac{a_N}{a_N+\lambda}) \end{bmatrix} \qquad (10)$$

Then we can rewrite the variance as $\mathrm{Var}[\hat{y}|X,\lambda] = \sigma_\epsilon^2 VM^2V^T$ and the standard deviation of $\hat{y}$ as $\sigma_\epsilon VMV^T$. In this form we can easily see that when $\lambda = 0$, $M = I$ and thus $\mathrm{Var}[\hat{y}|X,\lambda] = \sigma_\epsilon^2 I$, i.e., the variance never collapses regardless of N. A useful interpretation of this results is that without regularization ($\lambda = 0$), adding more observations serves only to increase the available complexity of the best-fitting function, rather than reducing uncertainty over the parameters. By contrast, choosing $\lambda > 0$ allows additional observations to translate at least partly into improved certainty rather than additional complexity.

We now examine the behavior of the diagonal elements of $\sigma_\epsilon^2 VM^2V^T$ by examining their sum (the trace of $\sigma_\epsilon^2 VM^2V^T$), and then by extension their average. First, note the equality:

$$\begin{aligned}\operatorname{tr}[\sigma_\epsilon^2 V M^2 V^T] &= \sigma^2 \operatorname{tr}[V M^2 V^T] \\ &= \sigma_\epsilon^2 \operatorname{tr}[V^T V M^2] \\ &= \sigma_\epsilon^2 \operatorname{tr}[M^2] \\ &= \sigma_\epsilon^2 \sum_i \left(\frac{a_i}{a_i + \lambda}\right)^2 \qquad (11)\end{aligned}$$

Thus the sum of the variances of $\hat{y}$ is simply $\sigma_\epsilon^2 \operatorname{tr}[M^2]$. In this way the quantity $\operatorname{tr}[M^2]$ is central to understanding the consistency of KRLS. For large eigenvalues, the corresponding diagonal element of M^2 approaches 1. For eigenvalues near 0 by contrast, the corresponding element of M^2 approaches 0. Thus $\operatorname{tr}[M^2]$ can be roughly understood as the count of large eigenvalues of K or the number of important dimensions of the data. Given the construction of K, the number of large eigenvalues will grow with N only initially, after which additional observations will increase the number of important dimensions or eigenvalues only very rarely or not at all. Thus, for $\lambda > 0$, $\operatorname{tr}[M^2]$ will typically stop growing, or grow only very slowly after a small number of initial observations.[1]

Recall that $\sigma_\epsilon^2 \operatorname{tr}[M^2]$ was shown to be the sum of $\operatorname{Var}[\hat{y}_i|X, \lambda]$ over all observations i. Since this quantity stops growing or grows only very slowly in N, the average variance of $\hat{y}_i$ must be decreasing. Thus, the first condition for consistency is that λ must be greater than zero. Second, $\operatorname{tr}[M^2]$ must grow less quickly than N, which occurs so long as not every new observation leads to a relatively unique column of K (and thus another large eigenvalue). For large enough N, $\operatorname{tr}[M^2]$ slows in growth and eventually converges to a constant, and the average variance of $\hat{y}_i$ is thus $\frac{\sigma_\epsilon^2}{N} \operatorname{tr}[M^2]$.[2]

Note that in general the "curse of dimensionality" is not too severe for KRLS as it is not a strictly local method. However, higher dimensional data are more costly in terms of the variance

[1] The finite number of large eigenvalues as N grows, and thus the limited size of $\operatorname{tr}[M]$ as N grows is not only an empirical regularity, but also a result of the construction of K and the choice of σ^2. The only way for new observations to generate a large new eigenvalue is for that observation to be very different from the existing ones, thus producing a unique column of K. However, since we standardize the data in X after seeing all the data (and choose σ^2 to ensure a reasonable average distance between exemplars), we effectively bound the number of ways in which observations can differ. Put differently, as N grows large, each new x_i is increasingly likely to be similar to one already observed and thus the column of K it creates will be similar to one already present, so the corresponding eigenvalue is increasingly likely to be small.

[2] The above analysis treats λ as fixed, and indeed the diminishing variance depends on λ not being reduced too fast as N grows. In the way it has been written throughout the paper, λ should not depend heavily on N. Indeed in practice, the λ chosen by cross-validation does not generally diminish with N (given a fixed σ^2).

and its rate of convergence. Even after re-scaling the similarity metric using the rule derived above for σ^2, higher-dimensional spaces allow for a greater number of ways in which exemplars can be different from each other. The matrix K will thus have more dissimilar columns (with lower degrees of linear dependence between them), and can support more large eigenvalues. Regularization still ensures that the model is not over-fit on this set of relatively less dependent basis functions; however, the cost is borne in the variance. Under high-dimensionality, N must grow much larger before the number of eigenvalues of K effectively stops growing. This prolonged growth of the number of large eigenvalues implies that $\text{tr}[M]$ continues to grow as well, and thus the average variance of each $\hat{y}$, $\frac{\sigma_\epsilon^2}{N}\,\text{tr}[M^2]$, will shrink slower than $\frac{1}{N}$ until N becomes sufficiently large.

8. Proof of Theorem 4: Normality in Finite Samples

Here we establish the normality of the KRLS estimator around its expectation at a given point. First, we establish that the estimator is normally distributed in finite samples when the elements of ϵ are i.i.d. normal.

Assumption 5 (Normality) *The errors are distributed normally, $\epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$.*

Theorem 4 states that under assumptions 1–5, $\hat{y} \sim N(y^\star, (\sigma_\epsilon K(K+\lambda I)^{-1})^2)$.
The proof is as follows. For observation i the difference between the predicted and expected estimate, $\hat{y}_i - y_i^\star$, is given by the $i'th$ element of $K(K+\lambda I)^{-1}\epsilon$. Taking the data as fixed, this quantity is simply a linear combination of the i.i.d. normal elements of ϵ, and is thus also normally distributed. Therefore, when ϵ is i.i.d. normal, $\hat{y}$ is normally distributed in finite samples, with mean $y^\star$ and standard error as given above, $\sigma_\epsilon K(K+\lambda I)^{-1}$.

9. Proof of Theorem 5: Asymptotic Normality

Next, we establish that the normality also holds asymptotically even when the distribution of the residuals is not known to be normal.

Assumption 6 (Regularity Condition II) *Let (i) the errors be independently drawn from a distribution with finite mean and variance and (ii) the standard Lindeberg conditions hold such that the sum of variances of each term in the summation $\sum_j [K(K+\lambda I)^{-1}]_{(i,j)}\epsilon_j$ goes to infinity as*

$N \to \infty$ *and that the summands are uniformly bounded, i.e. there exists some constant* a *such that* $|[K(K+\lambda I)^{-1}]_{(i,j)}\epsilon_j| \leq a$ *for all* j*.*

Theorem 5 states that under assumptions 1–4 and 6, $\hat{y} \xrightarrow{d} N(y^\star, (\sigma_\epsilon K(K+\lambda I)^{-1})^2)$ as $N \to \infty$. The proof is as follows. Consider the gap between the estimated and expected value of $\hat{y}$ for a given observation, $\hat{y}_i - y^\star_i = [K(K+\lambda I)^{-1}\epsilon]_i$. This quantity can be thought of as

$$\sum_j [K(K+\lambda I)^{-1}]_{(i,j)}\epsilon_j \tag{12}$$

where $[K(K+\lambda I)^{-1}]_{(i,j)}$ gives the i^{th}, j^{th} element of $K(K+\lambda I)^{-1}$. In this sum of weighted ϵ_j each element remains independent and has finite mean and variance. Under standard regularity conditions, this summated quantity is normally distributed by the Lindeberg-Feller central limit theorem.[3]

10. Comparing KRLS on Interpretability and Inference

As noted, this paper is designed to introduce KRLS to social scientists as a potential addition to their toolbox and to extend the method analytically and computationally to make it more effective for social science inquiry. Thus while we have compared the performance of KRLS to several other approaches in the text (e.g. Table 3) and found it performs very favorably, we are primarily interested in how interpretation and inference with KRLS compares to other useful and important approaches.

We begin with the supposition that most social science investigators, when interpreting their fitted data, seek measures that quantify the magnitude and uncertainty of the marginal effects of the input variables or functions thereof. Ideally, the output of such a procedure would appear similar to a regression table, with effect estimates and variances allowing for hypothesis tests or the construction of confidence intervals around estimated marginal effects. This requirement rules out many otherwise powerful approaches, including k-Nearest Neighbors (k-NN), neural networks (though marginal effects can generally be simulated, e.g. Beck et al. (2000)), Support Vector Machines (though their predictive performance is similar to KRLS, see Rifkin et al. (2003); Zhang

[3] The so-called Lindeberg condition is both sufficient and necessary for this result (Feller; 2008). In the regularity conditions cited here we have used a slightly stronger but more easily stated set of conditions (Grinstead and Snell; 1997).

and Peng (2004)), classification and regression trees (CART), random forests, or Bayesian Additive Regression Trees (BART), to name a few. While it may be possible in these cases to estimate partial derivatives with respect to each input variable and at each observation through repeated simulations, this would require extensive time, computation, and additional effort. If that procedure is not already difficult enough or computationally prohibitive, re-iterating that whole process hundreds or thousands of times to estimate bootstrap based variances of partial derivatives is even more likely to be infeasible. Even where possible, these procedures are certainly not convenient as a stepping stone for current GLM users.

Generalized Additive Models (GAMs, Hastie and Tibshirani (1990)) deserve further discussion as perhaps the best known non-GLM model among social science researchers. GAMs provide graphical plots of $E[y|x^{(j)}]$ for covariate j, from which a sense of $\frac{\partial y}{\partial x^{(j)}}$ for each covariate can be gleaned.[4] A major concern with GAMs is that while they are very useful in many cases and the underlying theory is well developed (see e.g. Wood; 2006), the additivity constraint is likely too restrictive for many social science problems: we often expect the marginal effect of one variable to change across levels of other variables (see simulations for examples). While GAMs can answer this concern by allowing some variables to be "smoothed together", this requires the user to know in advance what (small) set of variables should be smoothed together. Our general supposition is that the user does not have this information, and in many cases, should not be given the latitude to guess and choose among the most favorable results.[5] Nevertheless, in cases where the investigator has strong reason to suspect that most or all variables act additively, then GAMs are a good choice.

Remaining approaches that allow the marginal effects to be directly read off include linear models and expansions of them to include transforms of the independent variables to add flexibility. A promising approach in this family is to allow the user to implement a large number of expansions of the columns of the predictor matrix. The expansions may include polynomial functions of individual

[4] Though it is not convenient and we have never seen this done by applied researchers, some choices of basis functions for the smoothing of each separate covariate make it possible mathematically to compute partial derivatives, which could then be summarized as suggested here.

[5] For certain choices of multidimensional basis functions, and particularly the thin plate spline regression basis functions (which do not require selecting knots), fitting GAMs by penalized least squares over these bases is similar to conducting KRLS with a different kernel. However, while GAM uses a truncated set of basis functions to regress on, and has computational difficulties in smoothing over more than a few variables, KRLS with a thin-plate spline kernel would instead utilize inner-products of the associated expansions via the kernel trick. Moreover, appropriate interpretational machinery would also need to be added to current implementations to make it usable in ways similar to KRLS.

variables, multiplicative or tensor product interactions, piecewise constants or linear components (as in Multivariate Adaptive Regression splines), trigonometric functions, or other choices of feature mappings. Linear models can then be fit directly in the column space of these expanded bases, even when their dimension exceeds N, using a regularization approach such as ridge regression or "LASSO" (least absolute shrinkage selection operation) (Tibshirani; 1996). The LASSO is, like KRLS, a least squares, regularized model, but uses a penalty term that results in sparsity (i.e. it drives some of the coefficients to exactly zero rather than just shrinking them). The hope is that, if this selects a small number of coefficients to have non-zero values, interpretation follows, as these coefficients give the marginal effects of the corresponding bases.

This "explicit-expansion-then-LASSO" type approach has some advantages and may be a particularly useful alternative, especially for problems that require quickly choosing a (sparse) linear model from among a large number of input variables. However, it also has some drawbacks. First, the mapping from the input variables to the basis expansion must be done explicitly. This requires considerable guesswork,[6] and unfortunately, in practice LASSO is not generally stable to different choices of these expansions (the basis functions that survive often depend on which other functions are included). Related, if some input variables are highly correlated, the LASSO will often select one from the group, the choice of which may be unstable against small perturbations of the data. Moreover, while polynomials expansions and multiplicative interactions offer added flexibility, there is no reason to expect they are generally appropriate basis functions to use. As we have argued here the KRLS basis functions are motivated in terms of the "similarity view", ensuring its suitability to many problems in social science where smoothness (high correlation of nearby points, lesser correlation of distant points) is a minimal reliable assumption.[7]

In this sense the KRLS approach to modeling and interpretation is fundamentally different from the "explicit-expansion-then-LASSO" approach. In KRLS the goal is to "get the CEF right" first (using the mapping of X to K and then running a ridge regression which penalizes more complex

[6] For example, should first, second, third, or even higher order interactions of all independent variables be included? What order of polynomials should be included for the continuous covariates?

[7] A Taylor expansion view may justify the use of polynomial functions of the inputs as allowing it to model a generic smooth function of the inputs. However, this is applicable only in the neighborhood around the point of expansion. Typically, explicit polynomial expansions are taken only around 0. Introducing a lattice of "knots" throughout the covariate space and expanding around these would solve this (effectively by producing splines) but this adds considerably to the complexity and the difficulty of interpretation, and is rarely done in practice in the context of the expansion-then-LASSO approach.

functions), and then to read off the (possibly nonconstant) partial derivatives with respect to the original input variables X. This is straightforward, since KRLS guarantees that the model will map onto a set of pointwise partial derivatives and these marginal effects are readily interpretable as in a simple linear regression. This approach has the benefit of avoiding the need to make arbitrary decisions about the number and nature of terms to include in the explicit basis expansion of X, thus prohibiting the user from manipulating the results through such choices.

Second, while sparsity is often equated with interpretability, sparsity is neither sufficient nor necessary for interpretability. Suppose that - as can be quite common in practice - the LASSO produces non-zero coefficients for a series of terms that include second- or third-order interactions, but not the main effects. Or suppose it comes back with fourth- or higher-order interactions. Both are difficult to interpret. In addition, an interpretation of how substantial such interactions are requires also considering the densities of the variables involved—a consideration automatically accounted for in the KRLS summaries. LASSO approaches would also not have the benefits of KRLS in terms of protecting against extreme counterfactuals.

Finally, since LASSO does both selection and model fitting together, inference is considerably more complicated. Had a different sample been drawn (or the original sample re-sampled in a bootstrap), a different set of variables may have been (and oftentimes are) selected to have non-zero coefficients. This implies that inference must include uncertainty as to variable selection as well. Indeed, the limiting distribution of the LASSO estimator is fairly complicated and researchers typically use a bootstrap to estimate variances, despite the fact that standard bootstrap method for the lasso estimator are known to be inconsistent (Knight and Fu; 2000; Chatterjee and Lahiri; 2010). This contrasts with KRLS, where the variances for the marginal effects can be conveniently estimated in closed form and the estimator is asymptotically normal so confidence intervals can also be easily constructed.

11. Standard Errors for First-Difference Estimator

While the estimated marginal effects at the observed exemplars are always available in closed-form, averaging over these provides a poor measure for the marginal effect of binary predictors, $d \in \{0, 1\}$. This is because the marginal effect of going from $d = 0$ to $d = 1$ is not well characterized by $\frac{\partial y}{\partial d}$ measured at $d = 0$ and $d = 1$ alone. We thus characterize the marginal effect of binary predictors through first differences. Consider a dataset with a binary predictor, d, other predictors X, and outcome variable y. We train the model on the actual data, $(y,\, [X, d])_i$, obtaining $\hat{c}$ and the full matrix $\mathrm{Var}[\hat{c}]$. We now wish to compute the sample analog to the expected first-difference over the distribution of the data,

$$\overline{FD} = \frac{1}{N}\sum_j [\hat{y}|X = X_j, d = 1] - \frac{1}{N}\sum_j [\hat{y}|X = X_j, d = 0] \tag{13}$$

First construct two new test sets: X_1, in which X takes its natural values but d always equals 1; and X_0, in which always X takes on its natural values but $d = 0$ everywhere. From these we compute the test kernel matrix (measuring similarity of each of these test observations to the original training observations), K_1 corresponding to X_1 and K_0 corresponding to X_0.

Next, construct $\hat{FD}$, the vector of estimated first differences indexed by i. This can be computed easily:

$$\begin{aligned} \hat{FD} &= (\hat{y_i}|X_i, d = 1) - (\hat{y_i}|X_i, d = 0) \\ &= K_1\hat{c} - K_0\hat{c} \\ &= (K_1 - K_0)\hat{c} \\ &= M\hat{c} \end{aligned} \tag{14}$$

where $M = K_1 - K_0$. Now, let h be a $N \times 1$ vector where each element is $1/N$. Then the average first difference, $\overline{FD}$, is computed as $h^t M\hat{c}$. The variance of this quantity is computed from:

$$\mathrm{Var}[\overline{FD}] = h^t M\, \mathrm{Var}[\hat{c}] M^t h \tag{15}$$

This method produces accurate estimates of uncertainty in simulations. For example, suppose $X_1 \sim N(0, 1)$, $X_2 \sim bernoulli(.5)$, $\epsilon \sim N(0, 1)$, and $y = X_1 + X_2 + \epsilon$. The analytical standard errors (SEs) produced for data simulated this way very closely match the observed distribution of estimated average first differences over repeated samples, and produced accurate 95% coverage rates. Using these SEs to construct confidence intervals (under a normal approximation) unsurprisingly requires that N is large enough for the central limit theorem to apply. The confidence intervals produced by this method were roughly 10-15% too small when $N = 100$. By $N = 200$, however, there was no detectable bias. In an experiment with 1000 iterations at $N = 200$, the average of the analytical

SEs over these runs was very close to the standard deviation of point estimates, differing by only 3%. Moreover, the 95% confidence interval (using a normal approximation and the analytical SE) had a true coverage rate of 95.1%.

12. Additional Illustrations and Simulation Results

12.1. High and Low Frequency Functions

A key assumption motivating the use of regularization is that we prefer smoother, less complicated functions. There are several motivations for this. First, we argue that in social science research, we often believe that the conditional expectation function characterizing the data-generating process is relatively smooth. We would not often believe conditional expectation functions that vary erratically (with the possible exception of those produced by sharp discontinuities in laws or policies). Second, though we traditionally assess goodness-of-fit based on the predicted values at observed points, the generalizability of a relationship is determined by how it behaves between observations. It seems reasonable then to prefer functions that do not imply wild oscillations in the outcome value at points located in-between the observed (training) points. This can be achieved by preferring functions that are less "wiggly" as measured by some norm over the entire function.

Put another way, for most social science inquiry we think that "low-frequency" relationships – in which y cycles up and down fewer times across the range of a given x – are theoretically more plausible and useful than "high-frequency" relationships. We illustrate this in figure A.2. The dotted line shows a higher-frequency relationship between a predictor x and an outcome y, which we would not believe to be an accurate representation of reality. The more generalizable, stable, or "true" relationship we believe to underly these data would likely be a lower-frequency function, such as the solid line, fitted by KRLS.

12.2. Example of Fitting a Non-linear Function and its Derivatives

As an illustration, consider a simple case such as $y = 100 + 3x^4$, and its derivative, $\frac{\partial y}{\partial x} = 12x^3$. These two functions are shown in A.3. We draw $X \sim Unif(-4, 4)$ for three sample sizes 10, 50, and 100 and simulate observed outcomes using $y = 100 + 3x^4 + \varepsilon$ where $\varepsilon \sim N(0, 1)$. The resulting fits from the KRLS estimator (averaged across 500 simulations) are reported in the left subplots of Figure A.3. The fitted values from KRLS accurately reproduce both the target function (black dots) and its derivative (gray triangles) across all three sample sizes.

Note that no model specification search or specific functional form assumption is required. Users simply pass y and X to the estimator; σ^2 is determined by the rule described above and λ is found by cross-validation. The subplots on the right show the biased estimates from the OLS regression for comparison.

12.3. Modeling Common Interactions

KRLS is well suited to fit target functions that are non-additive and or involve more complex interactions as they may arise in social science research. We consider three types of functions: those with one "hill" and one "valley", two hills and two valleys, or three hills and three valleys. These functions, especially the first two, are often represented by "two-by-two" tables, and correspond to rather common scenarios in the social sciences where the effect of one variable changes or dissipates depending on the effect of another. For example, left wing parties might lead to good economic outcome when labor organizations are strong, but poor outcomes when labor is weak (Alvarez et al.; 1991).

In figures A.4–A.6 we simulate the one hill/one valley, two hills/two valleys, and three hills/three valleys examples. In each case we use 200 observations, $x_1, x_2 \sim Unif(0, 1)$, and noise given by $\varepsilon \sim N(0, .25)$. We then fit these data using KRLS, OLS, and GAMs. Results are averaged over 100 simulation. As summarized in table 2 in the text, KRLS outperforms GAMs and OLS on both in- and out-of-sample fits for all three functions.[8]

13. Additional Applied Example: Brambor et al. 2006

At the suggestion of an anonymous reviewer, we also applied KRLS to an empirical example previously cited in Brambor et al. (2006), a paper well known for its discussion of the appropriate uses of multiplicative interaction terms in linear models. The example used in that paper, drawn from Golder (2006), is a test of the "short-coattails" hypothesis, which states that "temporally-proximate presidential elections will reduce the effective number of legislative parties if and only if the number

[8] We note that interpretational concerns aside, GAMs could perform suitably on these tasks as well, if the user knows in advance which two variables to "smooth together", which allows a multi-dimensional spline to be fit similar to KRLS, though the behavior of such functions outside the support of the data differ from KRLS. However, our general supposition is that the user may not know in advance which variables to interact in this way and the approach of smoothing together various predictors typically runs into numerical problems as the number of predictors increases.

of presidential candidates is sufficiently low." That is, suppose *parties* is the number of legislative parties in a given country and *proximity* measures how closely in time legislative elections occur to presidential elections. Then the short-coattails hypothesis states that more recent presidential elections will reduce the number of legislative parties due to a coattail effect (i.e. $\frac{\partial parties}{\partial proximity}$ will be negative), but only when there are relatively few presidential candidates. In the presence of too many presidential candidates the effect is expected to be weak or even reversed (i.e. $\frac{\partial parties}{\partial proximity}$ should approach or exceed zero as the number of candidates grows). The modeling approach in Brambor et al. (2006) is based on a specification with a linear interaction term given by:

$$\begin{aligned} ElectoralParties = \beta_0 + \beta_1 Proximity + \beta_2 PresidentialCandidates + \\ \beta_3 (Proximity \cdot PresidentialCandidates) + \beta_4 Controls + \epsilon \end{aligned} \tag{16}$$

The marginal effects from this model are shown in the top panel of Figure A.7. The findings strongly supports the short-coattails" hypothesis: at low numbers of presidential candidates, $\frac{\partial parties}{\partial proximity}$ is negative and highly significant; at high levels it comes back towards zero. The maximum negative effect is observed when the effective number of presidential candidates is zero.

Note that this approach to thinking about and modeling interactions by including multiplicative terms is highly constrained. In this case, the specification only allows the marginal effect to vary according to $\frac{\partial parties}{\partial proximity} = \beta_1 + \beta_3 PresidentialCandidates$. The KRLS approach does not begin with such a specification, but instead fits a smooth model of minimal complexity which allows the marginal effect to vary across the covariate space. The pointwise estimates for $\frac{\partial parties}{\partial proximity}$ as estimated by KRLS are shown in the bottom panel portion of Figure A.7.

The results reflect some similarities with the original findings from the linear interaction model, but also important differences. When there are over two presidential candidates, the results of KRLS look strikingly like those from the linear interaction model. However, with two or fewer presidential candidates, the results differ markedly. The KRLS model suggest that the marginal effect $\frac{\partial parties}{\partial proximity}$ is close to zero when the number of presidential candidates is zero, and it becomes slightly more negative as we move to the cases with one or two presidential candidates. This is in stark contrast to the linear interaction model which implies that the maximum effect is obtained at zero presidential candidates.

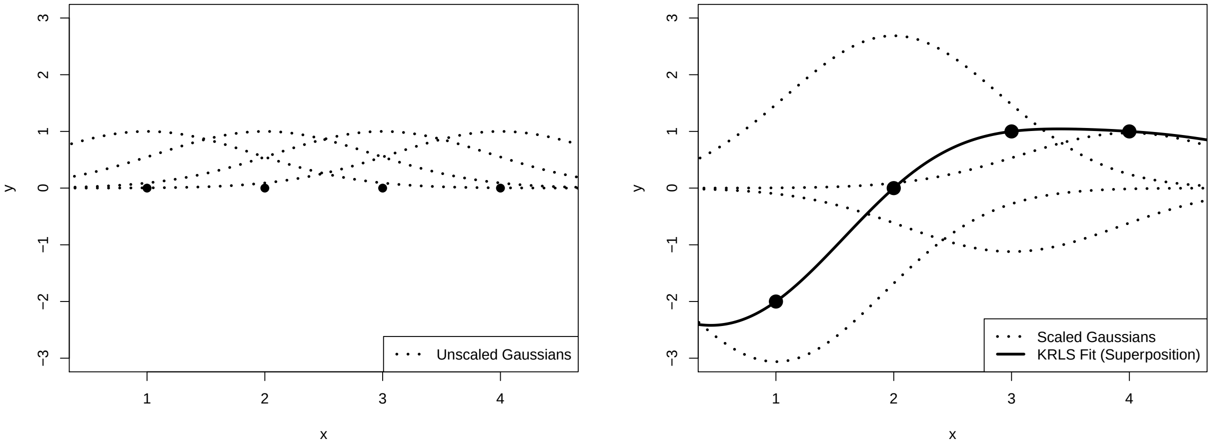
This difference in what the models reveal is important, given that 62% of the observations have zero presidential candidates. While we are not experts on the electoral systems studied in Golder (2006), the results also seem plausible: systems with zero one, or at most two presidential candidates may experience little or no coattail effect, as the number of legislative parties is likely to be largely unaffected by presidential campaigning or electoral results. Moreover, countries with zero presidential candidates are likely to be of a different type.[9] In this sample, as an anonymous reviewer noted, those with zero presidential candidates appear to be largely parliamentary democracies, making it somewhat awkward to pool them together with other countries having presidential elections in a single model. While problems related to the pooling of different types of units should perhaps be considered at an earlier level, KRLS proves valuable in revealing such problems by flexibly estimating heterogeneous marginal effects. Finally, with the insight suggested by the KRLS analysis, we can return to OLS analysis, but allow for the marginal effect to vary, by fitting different models to where there are two or fewer presidential candidates versus more than two. The results of such an analysis support the insight suggested by KRLS: at two or fewer candidates, $\frac{\partial parties}{\partial proximity}$ is close to zero and slightly decreasing in the number of candidates. At more than two candidates, $\frac{\partial parties}{\partial proximity}$ follows the pattern previously found: it is increasing in the number of candidates, reaching approximately zero when the number of candidates reaches its maximum. Accordingly, the coefficient on the interaction term is slightly negative (-0.12) when run on data with two or fewer presidential candidates, but when there are more than two, it is almost identical to the point estimate in the original finding (0.28, compared to original of 0.29). Figure A.8 shows the results graphically.

An important point here is that while adding multiplicative interaction terms to linear models only allows marginal effects to vary linearly, KRLS allows marginal effects to vary in virtually any smooth way – a difference that can be critical to the substantive inferences, as shown here. It is important to emphasize that we do not intend this replication to be a critique of Brambor et al. (2006). Instead, it again highlights that finding the correct functional form by adding interactions or higher order terms is difficult, and may not allow enough flexibility in the way marginal effects are allowed to change. We thank the reviewer for suggesting this interesting example.

[9] Golder notes that a country might have few presidential candidates for two reasons: (a) "the demand for presidential candidates is low because there are few social cleavages", and (b) "the demand for presidential candidates is high but the electoral system is not permissive."

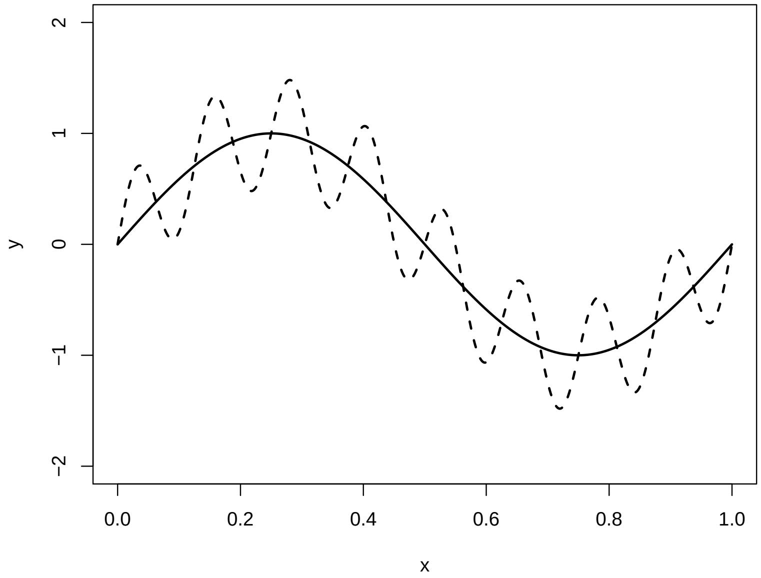
14. Additional Figures, Simulation Results

Figure A.1: Fitting a Simple Function with KRLS



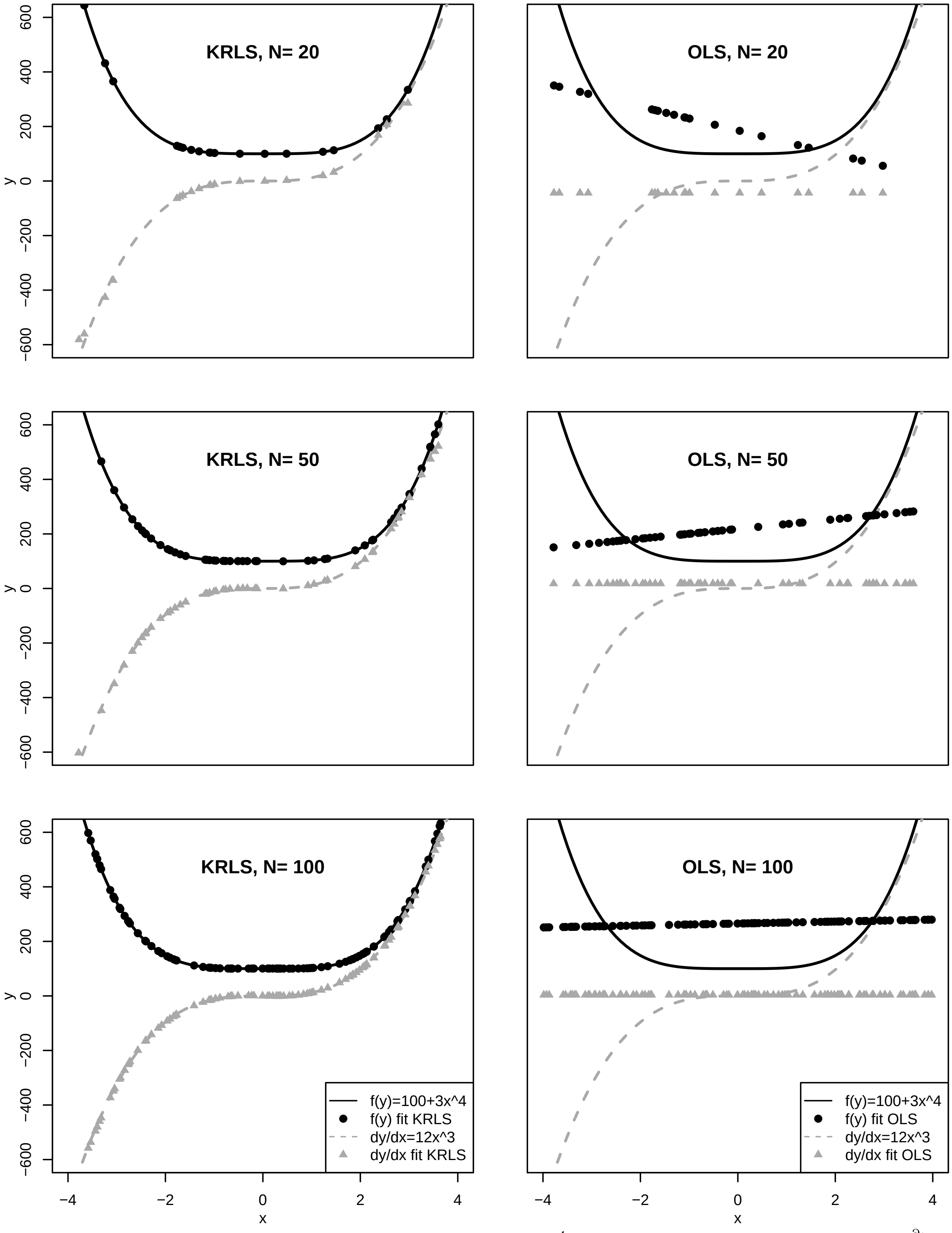
Note: Left Panel: Unscaled Gaussians placed over each of the four data points. Right Panel: Gaussians scaled by the choice coefficients obtained from KRLS. The choice coefficients for the data points (from left to right) are $c = [-3.06, 2.68, -1.12, 0.97]$

Figure A.2: Example of High- and Low-Frequency Functions



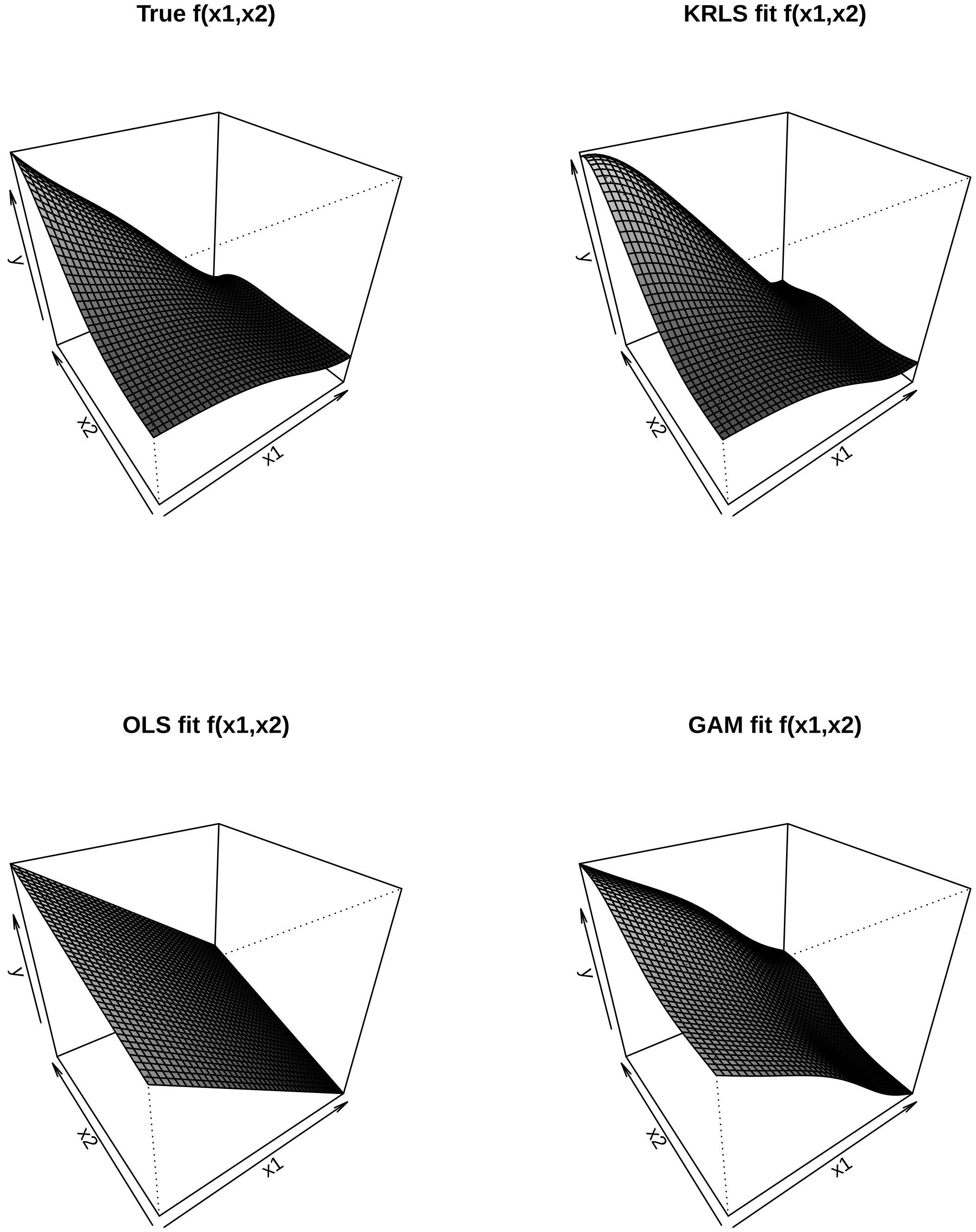
Note: The solid line represents a “good” explanation of the relationship between x and y. The dashed line represents a “bad” one, which is both considered more likely to be noise and is also much less useful in a theoretical way. For most social science inquiry, we are interested in recovering conditional expectation functions that look like the solid, low-frequency line, not the dashed, high-frequency line.

Figure A.3: KRLS Fits Non-Linear Functions and their Derivatives



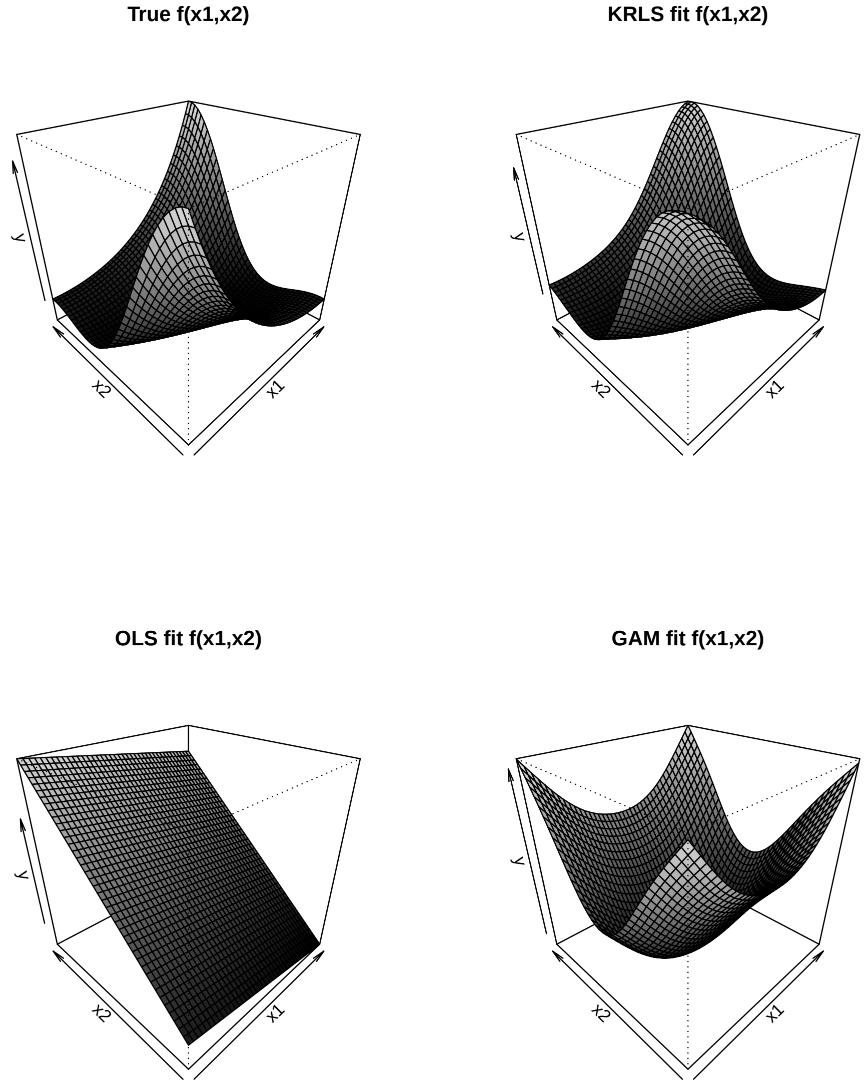
Note: Simulation to recover the non-linear function $y = 100 + 3x^4$ (black solid line) and its derivative $\frac{\partial y}{\partial x} = 12x^3$ (gray dashed line). The sample sizes are 10, 50, and 100, $X \sim Unif(-4, 4)$, and observed outcomes are simulated as $y = 100 + 3x^4 + \varepsilon$ where $\varepsilon \sim N(0, 1)$. In the left figures, the black dots show the fitted values for $\hat{y}$, and the grey triangles show the fitted values for $\hat{\frac{\partial y}{\partial x}}$ from the KRLS estimator (average across 500 simulations). The estimates in the right figures show the estimates from the OLS estimator accordingly.

Figure A.4: KRLS Approximates Complex Interactions: One Hill, One Valley



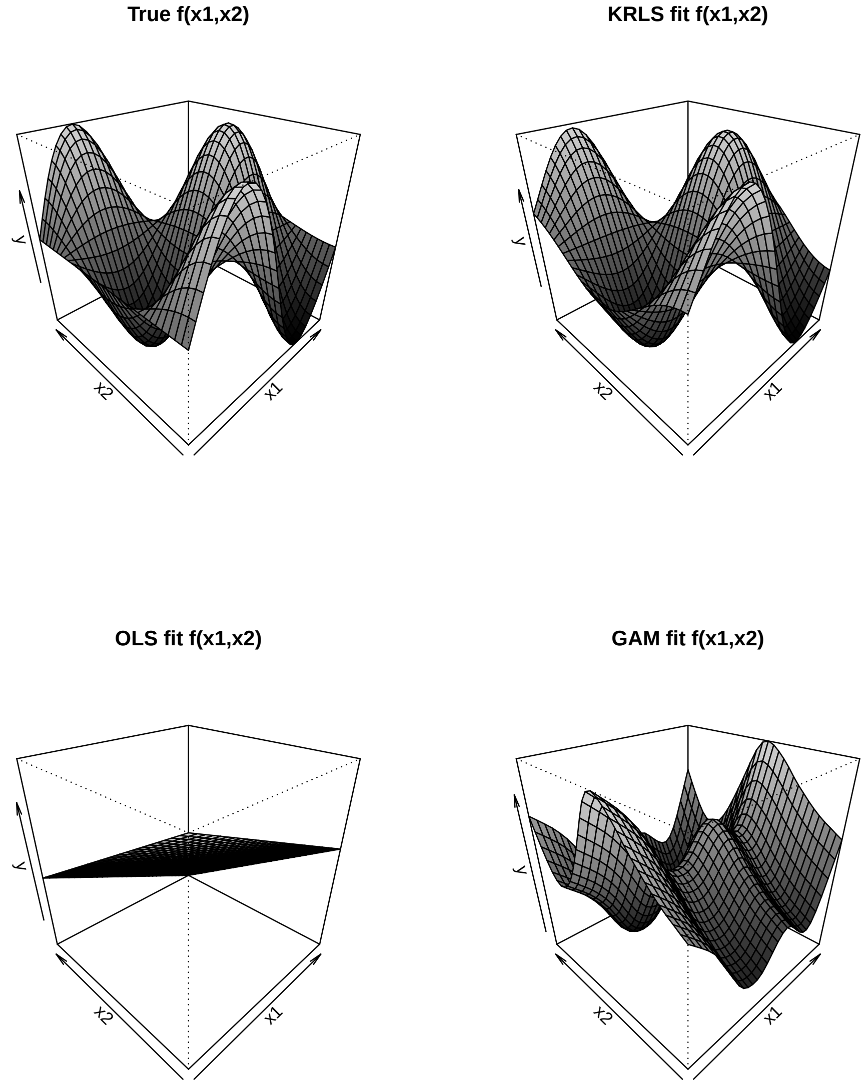
Note: Simulation to recover the target function given by $y = e^{-5(1-x_1)^2+(1-x_2)^2} - e^{-5(1-x_2)^2+(x_1)^2}$ using simulations with 200 observations drawn from $x_1, x_2 \sim Unif(0, 1)$ and random noise $\varepsilon \sim N(0, .25)$. The top right figure shows the true target function. The top left, bottom right, and bottom left figures show the fitted functions from the KRLS, OLS, and GAM respectively.

Figure A.5: KRLS Approximates Complex Interactions: Two Hills, Two Valleys



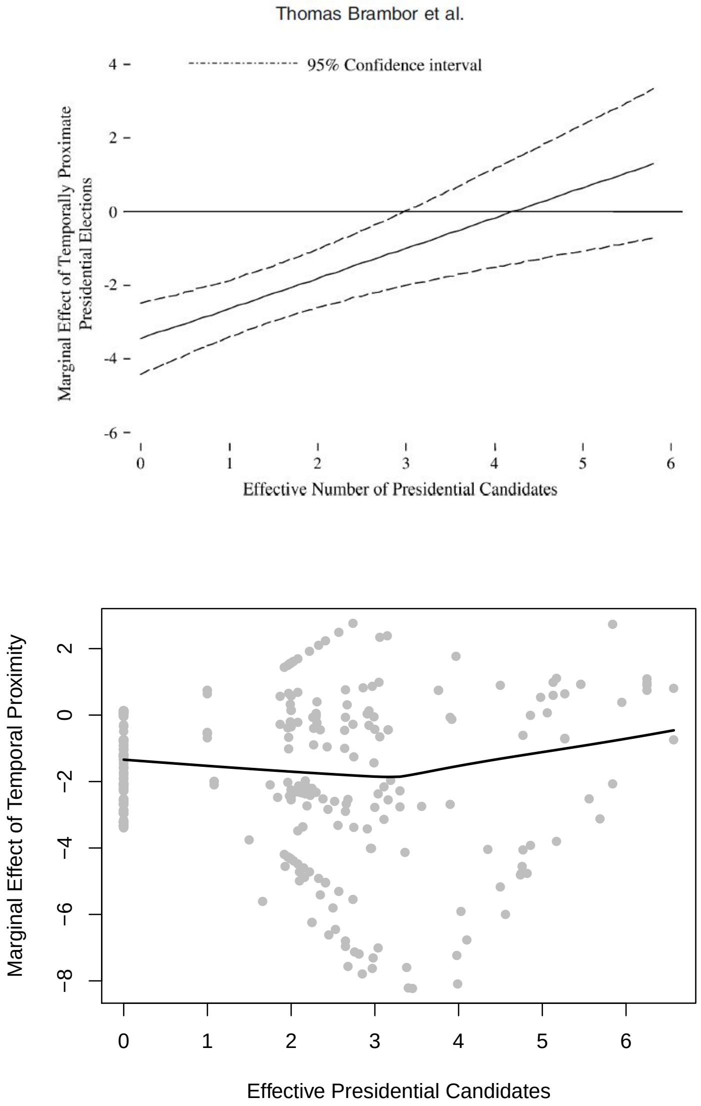
Note: Simulation to recover the target function given by $y = e^{-5(1-x_1)^2+(x_2)^2} + e^{-5(x_1)^2+(1-x_2)^2}$ using simulations with 200 observations drawn from $x_1, x_2 \sim Unif(0,1)$ and random noise $\varepsilon \sim N(0, .25)$. The top right figure shows the true target function. The top left, bottom right, and bottom left figures show the fitted functions from the KRLS, OLS, and GAM respectively.

Figure A.6: KRLS Approximates Complex Interactions: Three Hills, Three Valleys



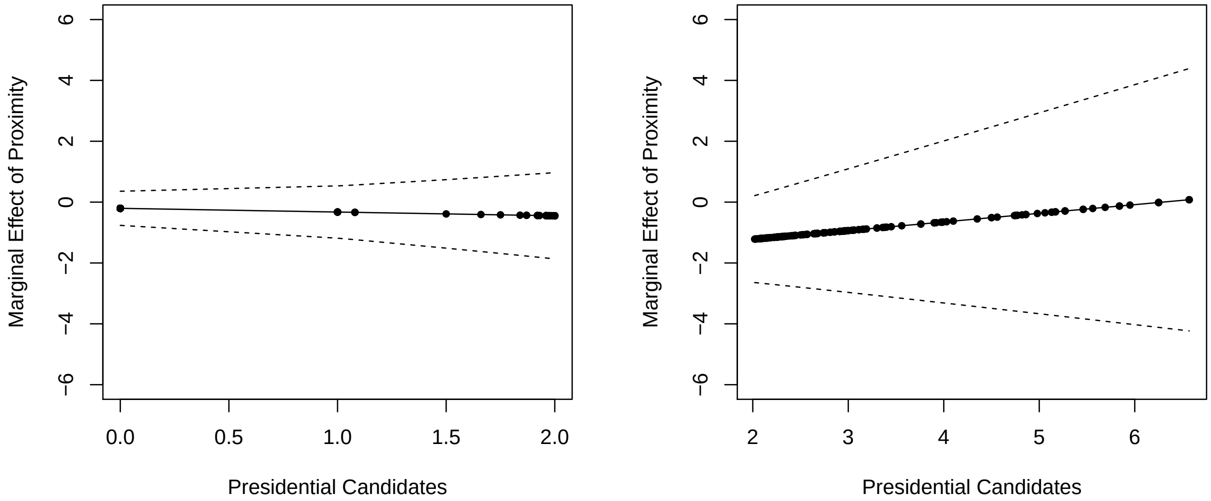
Note: Simulation to recover the target function given by $y = sin(x_1) * cos(x_2)$ using simulations with 200 observations drawn from $x_1, x_2 \sim Unif(0, 2\pi)$ and random noise $\varepsilon \sim N(0, .25)$. The top right figure shows the true target function. The top left, bottom right, and bottom left figures show the fitted functions from the KRLS, OLS, and GAM respectively.

Figure A.7: The marginal effect of temporally proximate presidential elections on the effective number of electoral parties



Note: Top Panel: Figure 3 from Brambor et al. (2006). More temporally proximate presidential and legislative elections lead to fewer effective electoral parties. However, this is true only when there are relatively few presidential candidates, and the effect vanishes when there are large numbers of presidential candidates. Bottom Panel: Scatterplot of pointwise marginal effects of temporal proximity on the number of parties ($\frac{\partial parties}{\partial proximity}$), with lowess estimates super-imposed. The plot looks similar to the Brambor et al. (2006) model only when there are 3 or more presidential candidates. By contrast, with zero presidential candidates (which represents 62% of the observations included in the Brambor et al. regression), the marginal effect estimates go back toward zero.

Figure A.8: OLS Results for Brambor et al. Split at Two Presidential Candidates



Note: Results from OLS models identical to those in the previous figure, but split at observations with two or fewer presidential candidates and those with more than two. The KRLS estimates differ from the original Brambor et al. (2006) results (A.7), suggesting that $\frac{\partial parties}{\partial proximity}$ takes values near zero when $PresidentialCandidates$ is zero (indicating no "coat-tail effect" there) and, if anything, decreases as $PresidentialCandidates$ rises to two and then reverses direction and follows the pattern suggested by Brambor et al. (2006) thereafter. Here, we split the sample and conduct OLS analyzes separately when $PresidentialCandidates \leq 2$ and when $PresidentialCandidates > 2$. As shown, the OLS results from the split samples reflect the KRLS results.

References

Alvarez, R., Garrett, G. and Lange, P. (1991). Government partisanship, labor organization, and macroeconomic performance, *The American Political Science Review* pp. 539–556.

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

Brambor, T., Clark, W. and Golder, M. (2006). Understanding interaction models: Improving empirical analyses, *Political Analysis* **14**(1): 63–82.

Chatterjee, A. and Lahiri, S. (2010). Asymptotic properties of the residual bootstrap for lasso estimators, *Proceedings of the American Mathematical Society* **138**(12): 4497–4509.

Feller, W. (2008). *An introduction to probability theory and its applications*, Vol. 2, Wiley-India.

Golder, M. (2006). Presidential coattails and legislative fragmentation, *American Journal of Political Science* **50**(1): 34–48.

Grinstead, C. and Snell, J. (1997). *Introduction to probability*, Amer Mathematical Society.

Hastie, T. and Tibshirani, R. (1990). *Generalized additive models*, Chapman & Hall/CRC.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators, *Annals of Statistics* pp. 1356–1378.

Rifkin, R., Yeo, G. and Poggio, T. (2003). Regularized least-squares classification, *Nato Science Series Sub Series III Computer and Systems Sciences* **190**: 131–154.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 267–288.

Wood, S. (2006). *Generalized additive models: an introduction with R*, Vol. 66, CRC Press.

Zhang, P. and Peng, J. (2004). Svm vs regularized least squares classification, *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, Vol. 1, IEEE, pp. 176–179.