Kernel Regularized Least Squares: Moving Beyond Linearity and Additivity Without Sacrificing Interpretability

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

First version: April 2012
This version: March 2013

Abstract

We propose the use of Kernel Regularized Least Squares (KRLS) for social science modeling and inference problems. KRLS borrows from machine learning methods designed to solve regression and classification problems without relying on linearity or additivity assumptions. The method constructs a flexible hypothesis space that uses kernels as radial basis functions and finds the best-fitting surface in this space by minimizing a complexity-penalized least squares problem. We argue that the method is well-suited for social science inquiry because it avoids strong parametric assumptions, yet allows interpretation in ways analogous to generalized linear models while also permitting more complex interpretation to examine non-linearities and heterogeneous effects. We also extend the method in several directions to make it more effective for social inquiry, by (1) deriving estimators for the pointwise marginal effects and their variances, (2) establishing unbiasedness, consistency, and asymptotic normality of the KRLS estimator under fairly general conditions, (3) proposing and justifying a simple automated rule for choosing the kernel bandwidth, and (4) providing companion software. We illustrate the use of the method through several simulations and a real-data example.

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.

Authors are listed in alphabetical order and contributed equally. We thank Jeremy Ferwerda, Dominik Hangartner, Lorenzo Rosasco, Marc Ratkovic, Teppei Yamamoto, and participants in seminars at NYU, MIT, the Midwest Political Science Conference, and the European Political Science Association Conference for helpful comments. Companion software written by the authors to implement the methods proposed in this paper in `R`, `Matlab`, and `Stata` can be downloaded at http://www.mit.edu/ jhainm/software.htm. The usual disclaimer applies.

1. Introduction

Generalized linear models (GLMs) remain the workhorse method for regression and classification problems in the social sciences. Applied researchers are attracted to GLMs because they are fairly easy to understand, implement, and interpret. However, GLMs also impose strict functional form assumptions. These assumptions are often unsustainable in social science data that are frequently ridden with non-linearities, non-additivity, heterogeneous marginal effects, complex interactions, bad leverage points, or other complications. It is well-known that misspecified models can lead to bias, inefficiency, incomplete conditioning on control variables, widely incorrect inference, and fragile model dependent results (e.g. King and Zeng (2006)). One traditional and well-studied approach to address some of these problems is to introduce high-order terms and interactions to GLMs (e.g. Friedrich 1982; Jackson 1991; Brambor et al. 2006). However, higher-order terms only allow for interactions of a prescribed type, and even for experienced researchers it is typically very difficult to choose among the many possible interaction specifications, which explode in number once the model involves more than a few variables. Moreover, as we show below, even when these efforts may appear to work based on model diagnostics, under common conditions they can instead make the problem worse, generating false inferences about the effects of included variables.

Presumably, many researchers are aware of these problems and routinely resort to GLMs not because they staunchly believe in the implied functional form assumptions, but because they lack convenient alternatives that relax these modeling assumptions while maintaining a high degree of interpretability. While some more flexible methods, such as neural networks (e.g. Beck et al. 2000) and Generalized Additive Models (GAMs, e.g. Wood 2003), have been proposed, they have not been widely adopted by social scientists, perhaps because these models often do not generate the desired quantities of interest or allow inference on them (e.g. confidence intervals or tests of null hypotheses) without non-trivial modifications and often impracticable computational demands.

In this paper, we build on a method rooted in machine learning approaches that we call Kernel Regularized Least Squares (KRLS). Our approach draws from Regularized Least Squares (RLS), a well-established method in the machine learning literature (see e.g. Rifkin et al. 2003).[1] We add the "K" to (a) emphasize that it employs kernels (whereas RLS could be applied to non-kernelized

[1] Similar methods appear under various names including Regularization Networks (e.g. Evgeniou et al. 2000) and Kernel Ridge Regression (e.g. Saunders et al. 1998).

models); and (b) to designate the specific set of choices we have made in this version of RLS, procedures we developed to remove all parameter specification from the investigator's hands and, most importantly, extensions we have added relating to interpretability. Our KRLS approach offers a versatile and convenient modeling tool that strikes a compromise between the highly constrained GLMs that many investigators rely on, and more flexible but less interpretable machine learning approaches. KRLS helps researchers to protect their inferences against misspecification bias and does not require them to give up many of the interpretative and statistical properties they value. As we show below, the KRLS estimator has desirable statistical properties under weak functional form assumptions and is nonetheless directly interpretable in ways similar to linear regression while also making much richer interpretations possible. The estimator yields pointwise estimates of partial derivatives that characterize the marginal effects of each independent variable at each data point in the covariate space. The researcher can examine the distribution of these pointwise estimates to learn about the heterogeneity in marginal effects, or average them in order to obtain an average partial derivative similar to a β coefficient from linear regression.

Because it marries flexibility with interpretability, the KRLS approach is applicable for a range of tasks, including exploratory analysis to learn about the data-generating process, model-based causal inference, or prediction problems that require an accurate approximation of a potentially highly non-linear conditional expectation function to impute missing counterfactuals. Similarly, it can be employed for propensity score estimation or other regression and classification problems where it is critical to use all the available information from covariates to estimate a quantity of interest. Instead of engaging in a tedious specification search, researchers simply pass the X matrix of predictors to the KRLS estimator (e.g. `krls(y=y,X=X)` in our R package), which then learns the target function from the data. For those who work with matching approaches, the KRLS estimator has the benefit of similarly weak functional form assumptions while allowing continuous valued treatments, maintaining good properties in high-dimensional spaces where matching and other local methods suffer the curse of dimensionality, and producing principled variance estimates in closed form. Finally, although necessarily less efficient than OLS, the KRLS estimator also has advantages even when the true data-generating process is linear, as it protects against model dependency that results from bad leverage points or extrapolation, and is designed to bound over-fitting.

The main contributions of this paper are threefold. First, we explain and justify the underlying methodology in an accessible way and introduce interpretations that illustrate why KRLS is a good fit for social science data. Second, we develop various analytical extensions. We (a) derive closed-form estimators for pointwise and average marginal effects; (b) derive closed-form variance estimators for these quantities to enable construction of confidence intervals; (c) establish the unbiasedness, consistency, and asymptotic normality of the estimator for fitted values under conditions more general than those required for GLMs; and (d) derive justification for a simple rule for choosing the bandwidth of the kernel at no computational cost, thereby taking all parameter-setting decisions out of the investigator's hands to improve falsifiability. Third, we provide companion software that allows researchers to implement the approach in `R`, `Stata`, and `Matlab`.

2. Explaining KRLS

Regularized least squares approaches with kernels, of which KRLS is a variant, can be motivated in a variety of ways. We begin with two explanations, the "similarity-based" view and the "superposition of Gaussians" view, which provide useful insight on how the methods work and why it is a good fit for many social science problems. We then present a third view that is perhaps less intuitive, but far more common in machine learning texts, to provide a more rigorous justification.[2]

2.1. Similarity-based View

Assume that we draw i.i.d. data of the form (y_i, x_i), where $i = 1, ..., N$ indexes units of observation, $y_i \in \mathbb{R}$ is the outcome of interest, and $x_i \in \mathbb{R}^D$ is our D-dimensional vector of covariate values for unit i (often called exemplars). Next, we need a so-called kernel, which for our purposes is defined as a symmetric and positive semi-definite function $k(\cdot, \cdot)$ taking two arguments and producing a real valued output.[3] It is useful to think of the kernel function as providing a measure of similarity between two input patterns. While many kernels are available, the kernel used in KRLS and throughout this paper is the Gaussian kernel given by

$$k(x_j, x_i) = e^{-\frac{||x_j - x_i||^2}{\sigma^2}} \tag{1}$$

[2] Another justification is based on the analysis of reproducing kernels, and the corresponding spaces of functions (Reproducing Kernel Hilbert Spaces) they generate along with norms over those spaces. For details on this approach, we direct readers to recent reviews included in Evgeniou et al. (2000) and Schölkopf and Smola (2002).

[3] By positive semi-definite, we mean that $\sum_i \sum_j \alpha_i \alpha_j k(x_i, x_j) \geq 0$, $\forall \alpha_i, \alpha_j \in \mathbb{R}, x \in \mathbb{R}^{\mathbb{D}}, D \in \mathbb{Z}^+$. Note that the use of kernels for regression in our context should not be confused with non-parametric methods commonly called "kernel regression" that involve using a kernel to construct a weighted local estimate.

where e^x is the exponential function and $||x_j - x_i||$ is the Euclidean distance between the covariate vectors x_j and x_i. This function is the same function as the normal distribution, but omits the normalizing factor $1/\sqrt{2\pi\sigma^2}$. The most important feature of this kernel is that it evaluates to its maximum of one only when $x_i = x_j$ and grows closer to zero as x_i and x_j become more distant. For present purposes, we will thus think of $k(x_i, x_j)$ as a measure of the *similarity* of x_i to x_j.

Under the "similarity-based view" we assert that the target function $y = f(x)$ can be approximated by some function in the space of functions represented by[4]

$$f(x) = \sum_{i=1}^{N} c_i k(x, x_i) \tag{2}$$

where $k(x, x_i)$ measures the similarity between our point of interest (x) and one of N input patterns x_i, and c_i is a weight for each input pattern. The key intuition behind this approach is that it does not model y_i as a linear function of x_i. Rather, it leverages information about the similarity between observations. To see this, consider some test point $x^\star$ at which we would like to evaluate the function value given fixed input patterns x_i and weights c_i. For such a test point, the predicted value is given by

$$\begin{aligned} f(x^\star) &= c_1 k(x^\star, x_1) + c_2 k(x^\star, x_2) + \ldots + c_N k(x^\star, x_N) & (3) \\ &= c_1(\text{similarity of } x^\star \text{ to } x_1) + c_2(\text{sim. of } x^\star \text{ to } x_2) + \ldots + c_N(\text{sim. of } x^\star \text{ to } x_N). & (4) \end{aligned}$$

That is, the outcome is linear in the similarities of the target point to each observation, and the closer $x^\star$ comes to some x_j, the greater the "influence" of x_j on the predicted $f(x^\star)$. This approach to understanding how equation (2) fits complex functions is what we refer to as the "similarity view." It highlights a fundamental difference between KRLS and the GLM approach. With GLMs we assume that the outcome is a weighted sum of the independent variables. In contrast, KRLS is based on the premise that there is information encoded in the similarity between observations, with more similar observations expected to have more similar outcomes. We argue that this latter approach is more natural and powerful in most social science circumstances: in most reasonable cases we do expect that the nearness of a given observation, x_i, to other observations reveals

[4] Here, we simply assert that the function of interest lies in this space, and we then show why this might be a suitable space of functions for many social science problems. We later provide a more conventional justification for this space based on ridge regressions in high-dimensional feature spaces.

information about the expected value of y_i, which suggests a large space of smooth functions in which observations close to each other in X are close to each other in y.

2.1.1. Superposition of Gaussians View

Another useful perspective is the "superposition of Gaussians" view. Recalling that $k(\cdot, x_i)$ traces out a Gaussian curve centered over x_i, we slightly rewrite our function approximation as

$$f(\cdot) = c_1 k(\cdot, x_1) + c_2 k(\cdot, x_2) + \ldots + c_N k(\cdot, x_N). \tag{5}$$

The resulting function can be thought of as the superposition of Gaussian curves, centered over the exemplars (x_i) and scaled by their weights (c_i). Figure 1 illustrates six random samples of functions in this space. We draw 8 data points $x_i \sim Uniform(0,1)$ and weights $c_i \sim N(0,1)$ and compute the target function by centering a Gaussian over each x_i, scaling each by its c_i, and then summing them (the dots represent the data points, the dotted lines refer to the scaled Gaussian kernels, and the solid lines represent the target function created from the superposition).[5] This figure shows that the function space is much more flexible than the function spaces available to linear regression; it enables us to approximate highly non-linear and non-additive functions that may characterize the data-generating process in social science data. The same logic generalizes seamlessly to multiple dimensions.

In this view, for a given realized dataset, KRLS would fit the target function by placing Gaussians over each of the observed exemplars x_i, and scaling them such that the summated surface approximates the target function. The process of fitting the function requires solving for the N values of the weights c_i. We, therefore, often refer to the c_i weights as *choice coefficients*, similar to the role that β coefficients play in linear regression. In reality, a great many choices of c_i can produce highly similar fits – a problem resolved in the next section through regularization. (In the web appendix we present a toy example to build intuition for the mechanics of fitting the function, see Figure A.1).

Before describing how KRLS chooses the choice coefficients, we introduce a more convenient matrix notation. Let matrix K be the $N \times N$ symmetric *Kernel matrix* whose j^{th}, i^{th} entry is $k(x_j, x_i)$; it measures the pairwise similarities between each of the N input patterns x_i. Let

[5] The center of the Gaussian curves depends on the point x_i, its upwards or downward direction depends on the sign of the weight c_i, and its amplitude depends on the magnitude of the weight c_i (as well as the fixed σ^2).

$c = [c_1, ..., c_N]^T$ be the $N \times 1$ vector of choice coefficients and $y = [y_1, ..., y_N]^T$ be the $N \times 1$ vector of outcome values. Equation (2) can be rewritten as

$$y = Kc = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \dots & k(x_1, x_N) \\ k(x_2, x_1) & \ddots & & \\ \vdots & & & \\ k(x_N, x_1) & & & k(x_N, x_N) \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \\ c_N \end{bmatrix}. \tag{6}$$

In this form, we plainly see KRLS as fitting a simple linear model: we fit y for some x_i as a linear combination of basis functions or regressors, each of which is a measure of x_i's similarity to another observation in the dataset. Notice that the matrix K will be symmetric and positive semi-definite and, thus, invertible.[6] Therefore, there exists a "perfect" solution to the linear system $y = Kc$, or equivalently, there exists a target surface that is created from the superposition of scaled Gaussians that provides a perfect fit to each data point.

2.2. Regularization and the KRLS Solution

While extremely flexible, fitting functions by the method described above produces a perfect fit of the data, and invariably leads to over-fitting. This issue speaks to the ill-posedness of the problem of simply fitting the observed data: there are many solutions that are similarly good fits. We need to make two additional assumptions that specify which type of solutions we prefer. Our first assumption is that we prefer functions that minimize squared loss, which ensures that the resulting function has a clear interpretation as a conditional expectation function (of y conditional on x).

The second assumption is that we prefer smoother, less complicated functions. Rather than simply choosing c as $c = K^{-1}y$, we instead solve a different problem that explicitly takes into account our preference for smoothness and concerns for over-fitting. This is based on a common but perhaps under-utilized assumption: in social science contexts, we often believe that the conditional expectation function characterizing the data-generating process is relatively smooth, and that less "wiggly" functions are more likely to be due to real underlying relationships rather than noise. Less "wiggly" functions also provide more stable predictions at values in-between those in the original data.

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 x) are theoretically more plausible

[6] This holds as long as no input pattern is repeated exactly. We relax this in the following section.

and useful than “high-frequency” relationships. (Figure A.2 in the appendix provides an example for a low- and high-frequency explanation of the relationship between x and y.)[7]

To give preference to smoother, less complicated functions, we change the optimization problem from one that considers only model fit to one that also considers complexity. Tikhonov regularization (Tychonoff 1963) proposes that we search over some space of possible functions and choose the best function according to the rule

$$\underset{f \in H}{\operatorname{argmin}} \sum_i (V(f(x_i), y_i)) + \lambda \mathcal{R}(f) \tag{7}$$

where $V(y_i, f(x_i))$ is a loss function that computes how “wrong” the function is at each observation, $\mathcal{R}$ is a “regularizer” measuring the “complexity” of function f, and $\lambda \in \mathbb{R}^+$ is a scalar parameter that governs the tradeoff between model fit and complexity. Tikhonov regularization forces us to choose a function that minimizes a weighted combination of empirical error and complexity. Larger values of λ result in a larger penalty for the complexity of the function and higher priority for model fit; lower values of λ will have the opposite effect. Our hypothesis space, H, is the flexible space of functions in the span of kernels built on N input patterns or, more formally, the Reproducing Kernel Hilbert Spaces (RKHS) of functions associated with a particular choice of kernel.

For our particular purposes, we choose the regularizer to be the square of the L_2 norm, $\langle f, f \rangle_H = ||f||_K^2$ in the RKHS associated with our kernel. It can be shown that, for the Gaussian kernel, this choice of norm imposes an increasingly high penalty on higher-frequency components of f. We also always use squared-loss for V. The resulting Tikhonov regularization problem is, thus, given by

$$\underset{f \in H}{\operatorname{argmin}} \sum_i (f(x_i) - y_i)^2 + \lambda ||f||_K^2. \tag{8}$$

Tikhonov regularization may seem a natural objective function given our preference for low-complexity functions. As we show in the appendix, it also results more formally from encoding our prior beliefs that desirable functions tend to be less complicated and then solving for the most likely model given this preference and the observed data.

[7] This smoothness prior may prove wrong if there are truly sharp thresholds or discontinuities in the phenomenon of interest. Rarely, however, is a threshold so sharp that it cannot be fit well by a smooth curve. Moreover, most political science data has a degree of measurement error. Given measurement error (on x), then, even if the relationship between the “true” x and y was a step function, the observed relationship with noise will be the convolution of a step function with the distribution of the noise, producing a smoother curve (for example, a sigmoidal curve in the case of normally distributed noise).

To solve this problem, we first substitute $f(x) = Kc$ to approximate $f(x)$ in our hypothesis space H.[8] In addition, we use as the regularizer the norm $||f||_K^2 = \sum_i \sum_j c_i c_j k(x_i, x_j) = c^T Kc$. The justification for this form is given below; however, a suitable intuition is that it is akin to the sum of the squared c_i's, which itself is a possible measure of complexity, but weighted to reflect overlap that occurs for points nearer to each other. The resulting Tikhonov problem is

$$c^\star = \underset{c \in \mathbb{R}^D}{\operatorname{argmin}}\, (y - Kc)^T (y - Kc) + \lambda c^T Kc. \tag{9}$$

Accordingly, $y^\star = Kc^\star$ provides the best-fitting approximation to the conditional expectation of the outcome in the available space of functions, given regularization. Notice that this minimization is equivalent to a ridge regression in a new set of features, one that measures the similarity of an exemplar to each of the other exemplars. As we show in the appendix, we explicitly solve for the solution by differentiating the objective function with respect to the choice coefficients c and solving the resulting first-order conditions, finding the solution $c^\star = (K + \lambda I)^{-1} y$.

We, therefore, have a closed-form solution for the estimator of the choice coefficients that produces the solution to the Tikhonov regularization problem within our flexible space of functions. This estimator is numerically rather benign. Given a fixed valued for λ, we compute the kernel matrix and add λ to its diagonal. The resulting matrix, is symmetric and positive definite and so inverting it is straightforward. Also, note that the addition of λ along the diagonal ensures that the matrix is well-conditioned (for large enough λ), which is another way of conceptualizing the stability gains achieved by regularization.

2.3. Derivation from an Infinite-Dimensional Linear Model

The above interpretations motivate the choices made in KRLS through our expectation that "similarity matters" more than linearity and that within a broad space of smooth functions, less complex functions are preferable. Another important route to this approach offers perhaps less intuition, but has the benefit of being generalizable to other choices of kernels and justifying both the choice of $f(x_i) = \sum_{j=1}^N c_j k(x_i, x_j)$ for the function space and $c^T Kc$ for the regularizer. For any positive semi-definite kernel function $k(\cdot, \cdot)$, there exists a mapping $\phi(x)$ that transforms x_i to a higher-dimensional vector $\phi(x_i)$, such that $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$. In the case of the Gaussian kernel,

[8] As we explain below, we do not need an intercept since we work with demeaned data for fitting the function.

the mapping $\phi(x_i)$ is infinite-dimensional. Suppose we wish to fit a regularized linear model (i.e. a ridge regression) in the expanded features, i.e. $f(x_i) = \phi(x_i)^T\theta$, where $\phi(x)$ has dimension D' (which is ∞ in the Gaussian case), and θ is a D' vector of coefficients. Then, we solve:

$$\underset{\theta \in \mathbb{R}^{D'}}{\operatorname{argmin}} \sum_i (y_i - \phi(x_i)^T\theta)^2 + \lambda||\theta||^2 \tag{10}$$

where $\theta \in \mathbb{R}^{D'}$ gives the coefficients for each dimension of the new feature space, and $||\theta||^2 = \theta^T\theta$ is simply the L_2 norm in that space. The first-order condition is $-2\sum_i^N (y_i - \phi(x_i)^T\theta)\phi(x_i) + 2\lambda\theta = 0$. Solving partially for θ gives $\theta = \lambda^{-1}\sum_{i=1}^N (y_i - \phi(x_i)^T\theta)\phi(x_i)$, or simply

$$\theta = \sum_{i=1}^N c_i\phi(x_i) \tag{11}$$

where $c_i = \lambda^{-1}(y_i - \phi(x_i)^T\theta)$. Equation 11 asserts that the solution for θ is in the span of the features, $\phi(x_i)$. Moreover, it makes clear that the solution to our potentially infinite-dimensional problem can be found in just N parameters, and using only the features at the observations.[9]

Substituting θ back into $f(x) = \phi(x)^T\theta$, we get

$$f(x) = \sum_{j=1}^N c_i\phi(x_i)^T\phi(x) = \sum_i^N c_i k(x, x_i) \tag{12}$$

which is precisely the form of the function space we previously asserted. Note that the use of kernels to compute inner-products between each $\phi(x_i)$ and $\phi(x_j)$ in equation 12 prevents us from needing to ever explicitly perform the expansion implied by $\phi(x_i)$; this is often referred to as the kernel "trick" or kernel substitution. Finally, the norm in equation 10, $||\theta||^2$ is $\langle\theta, \theta\rangle = \langle\sum_{i=1}^N c_i\phi(x_i), \sum_{i=1}^N c_i\phi(x_i)\rangle = c^TKc$. Thus, both the choice of function space and norm can be derived from a ridge regression in high- or infinite-dimensional feature space $\phi(x)$ associated with the kernel.

3. KRLS in Practice: Parameters and Quantities of Interest

In this section, we address some remaining features about the KRLS approach and discuss the quantities of interest that can be computed from the KRLS model.

[9] This powerful result is more directly shown by the Representer theorem (Kimeldorf and Wahba 1970).

3.1. Why Gaussians?

While users can build a kernel of their choosing to be used with KRLS, the logic is most applicable to kernels that radially measure the distance between points. We seek functions $k(x_i, x_j)$ that approach 1 as x_i and x_j become identical and approach 0 as they move far away from each other, with some smooth transition in between. Among kernels with this property, Gaussian kernels provide a sensible choice. One intuition for this is that we can imagine some data-generating process that produces x's with normally distributed errors. Some x's may be essentially "the same" point, but separated in observation by random fluctuations. Then, the value of $k(x_i, x_j)$ is proportional to the likelihood of the two observations x_i and x_j being the "same" in this sense. Moreover, we can take derivatives of the Gaussian kernel and, thus, of the response surface itself, which is central to interpretation.[10]

3.2. Data Pre-processing

We standardize all variables prior to analysis by subtracting off the sample means and dividing by the sample standard deviations. Subtracting off the means of the dependent variable is uncontroversial: it is equivalent to including an (unpenalized) intercept, but simplifies the mathematics and exposition. Subtracting the means of the x's has no effect, as the kernel is translation-invariant. The re-scaling operation is more controversial, though also commonly invoked in other penalized regressions. We make this choice for two reasons. First, re-scaling makes the methods invariant to unit-of-measure decisions. This is desirable since we have no reason to believe that measuring something in feet instead of inches, for example, should change the result. Second, re-scaling enables us to use a simple and fast approach for choosing σ^2, described next.

3.3. Choosing the kernel bandwidth σ^2

The choice of σ^2 governs how distant two input vectors x_i and x_j can be from each other and still be relatively similar. In the machine learning literature, σ^2 is often chosen empirically by some cross-validation technique. Here, we aim to minimize the number of parameters chosen by cross-validation or by the user, and instead rely on priors about typical social science data. A reasonable

[10] In addition, by choosing the Gaussian kernel KRLS is made similar to Gaussian Process regression, in which each point (y_i) is assumed to be a normally distributed random variable, and part of a joint normal distribution together with all other y_j, with the covariance between any two observations y_i, y_j (taken over the space of possible functions) being equal to $k(x_i, x_j)$.

requirement for social science data is that at least some observations can be considered similar to each other, some are different from each other, and many fall in-between. We propose setting $\sigma^2 = D$, where $D = dim(X)$ to generally satisfy this prior. A theoretical justification for this choice is that, for standardized data, the average (Euclidian) distance between two observations that enters into the kernel calculation, $E[||x_j - x_i||^2]$, is equal to $2D$ (see appendix). Choosing σ^2 by such a coarse rule of thumb may sacrifice some performance compared to agnostically optimizing σ^2 by an empirical method. However, our proposition is that any lost efficiency is typically offset by the simplicity and computational advantages of this approach. Additionally, our companion software allows the user to apply her own σ^2, and this feature can be used to implement more complicated approaches.

3.4. Choosing λ

Finally, we must choose the regularization parameter λ. The starting point for this is leave-one-out validation (LOOV): a model is trained on $N-1$ observations and tested on the left-out observation. This can be done N times for a given model with a specific choice of λ, and the errors on the left-out points can be averaged, giving the leave-one-out error for that choice of λ. A variant on this approach, generalized cross-validation (GCV), is equivalent to a weighted version of this loss (Golub et al. 1979). More importantly, we can compute the GCV error for any test value of λ without re-computing the choice coefficients each time, allowing us to compute this error in $O(N^1)$ time after one initial run (Rifkin and Lippert 2007).

4. Inference and Interpretation with KRLS

In this section, we establish properties of the KRLS estimator and develop quantities of interest.

4.1. Unbiasedness, Variance, Consistency, and Asymptotic Normality

We establish the unbiasedness, consistency, and asymptotic normality of the KRLS estimator. While statisticians and econometricians are often interested in these classical statistical properties, machine learning theorists have largely focused attention on whether and how fast the empirical error rate of the estimator converges to the true error rate. We are not aware of existing arguments for unbiasedness, or the normality of KRLS point-estimates, though proofs of consistency, distinct from our own, have been given, including in frameworks with stochastic X (e.g. De Vito et al. 2005).

4.1.1. Unbiasedness

We first show that KRLS unbiasedly estimates the best approximation to the true conditional expectation function that falls in the available space of functions given on our preference for less complex 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, $y_{obs} = y + \epsilon$.*

These two conditions together constitute the "correct specification" requirement for KRLS. Notice that these requirements are analogous to the familiar correct specification assumption for the linear regression model, which states that the data-generating process is given by $y = x'\beta + \epsilon$. However, compared to linear regression or GLMs more generally, the functional form assumption in KRLS is much more flexible.

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.*

This assumption is equivalent to the usual zero conditional mean assumption used to establish unbiasedness for linear regression or GLMs more generally. Under these assumptions, we can establish the unbiasedness of the KRLS estimator, meaning that the expectation of the estimator for the choice coefficients that minimize the penalized least squares $\hat{c}^\star$ obtained from running KRLS on y_{obs} equals its true population estimand, $c^\star$. Given these unbiasedness results, we can also establish unbiasedness for the fitted values.

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

Theorem 2 (Unbiasedness of fitted values) *Under assumptions 1–2, $E[\hat{y}] = y^\star$.*
The proof is given in the appendix.

We emphasize that this definition of unbiasedness says only that the estimator is unbiased for the *best approximation to the conditional expectation function given penalization.*[11] In other words, unbiasedness here establishes that we get the correct answer in expectation for $y^\star$ (not y), regardless of noise added to the observations. While this may seem like a somewhat dissatisfying notion of unbiasedness, it is precisely the sense in which many other approaches including OLS are unbiased. If, for example, the "true" data-generating process includes a sharp discontinuity that we do not have a dummy variable for, then KRLS will always instead choose a function that smooths this out somewhat, regardless of N, just as a linear model will not correctly fit a non-linear function. The benefit of KRLS over GLMs is that the space of allowable functions is much larger, making the "correct specification" assumption much weaker.

4.1.2. Variance

Here, we derive a closed-form estimator for the variance of the KRLS estimator of the choice coefficients that minimizes the penalized least squares, $c^\star$. This is important because it allows researchers to conduct hypothesis tests and construct confidence intervals. We utilize a standard homoscedasticity assumption, although the results could be extended to allow for heteroscedastic, serially correlated, or grouped error structures.[12]

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 1–3, the variance of the choice coefficients is given by* $\mathrm{Var}[\hat{c}^\star|K] = \sigma_\epsilon^2(K + \lambda I)^{-2}$. *The proof is given in the appendix.*

Lemma 2 (Variance of fitted values) *Under assumptions 1–3, the variance of the fitted values* $\hat{y}$ *is given by* $\mathrm{Var}[\hat{y}|X] = \mathrm{Var}[K\hat{c}^\star|X] = K^T[\sigma_\epsilon^2 I(K + \lambda I)^{-2}]K$.

[11] Readers will recognize that classical ridge regression (usually in the span of X rather than $\phi(X)$ or roughly speaking in the columns of K as here) is biased, in that the coefficients achieved are biased relative to the unpenalized coefficients. Imposing this bias is in some sense the purpose of ridge regression. However, if one is seeking to estimate the post-penalization function because regularization is desirable to identify the most reliable function for making new predictions, the procedure is unbiased for estimating that post-penalization function.

[12] We are currently working on implementing robust and cluster-robust standard errors in the companion software to allow researchers to further relax the homoscedasticity assumption.

In many applications, we also need to estimate the variance of fitted values for new counterfactual predictions at specific test points. We can compute these out-of-sample predictions using $\hat{y}_{test} = K_{test}\hat{c}^{\star}$ where K_{test} is the $N_{test} \times N_{train}$ dimensional kernel matrix that contains the similarity measures of each test observation to each training observation.[13]

Lemma 3 (Variance for test points) *Under assumptions 1–3, the variance for predicted outcomes at test points is given by* $\text{Var}[\hat{y}_{test}|X] = K_{test}\,\text{Var}[\hat{c}^{\star}|X]K_{test}^{T} = K_{test}[\sigma_{\epsilon}^{2}I(K+\lambda I)^{-2}]K_{test}^{T}$.

Our companion software implements these variance estimators. We estimate σ_{ϵ}^{2} by $\hat{\sigma}_{\epsilon}^{2} = \frac{1}{N}\sum_{i}^{N}\hat{\epsilon}^{2} = \frac{1}{N}(y - K\hat{c}^{\star})^{T}(y - K\hat{c}^{\star})$.

4.1.3. Consistency

In machine learning, attention is usually given to bounds on the error rate of a given method, and to how this error rate changes with the sample size. When the probability limit of the sample error rate will reach the irreducible approximation error (i.e. the best error rate possible for a given problem and a given learning machine), the approach is said to be consistent (e.g. De Vito et al. 2005). Here, we are instead interested in consistency in the classical sense, i.e. determining if $\underset{N\to\infty}{\text{plim}}\,\hat{y}_{i,N} = y_{i}^{\star}$ for all i. Since we have already established that $E[\hat{y}_{i}] = y_{i}^{\star}$, all that remains to prove consistency is that the variance of $\hat{y}_{i}$ goes to zero as N grows large.

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* $M < \infty$*.*

Theorem 3 (Consistency) *Under assumptions 1-4,* $E[\hat{y}_{i}|X] = y_{i}^{\star}$ *and* $\underset{N\to\infty}{\text{plim}}\,\text{Var}[\hat{y}|X] = 0$ *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 provided in the appendix.

Our proof provides several insights, which we briefly highlight here. The degrees of freedom of the model can be related to the effective number of non-zero eigenvalues. The number of effective eigenvalues, in turn, is given by $\sum_{i}\frac{a_{i}}{a_{i}+\lambda}$ where a_{i} are the eigenvalues of K. This generates two important insights. First, some regularization is needed ($\lambda > 0$), or else this quantity grows exactly

[13] To reduce notation, here we condition simply on X, but we intend this X to include both the original training data (used to form K) and the test data (needed to form K_{test}).

as N does. Without regularization ($\lambda = 0$), new observations translate into added complexity rather than added certainty; and accordingly the variances do not shrink. Thus, consistency is achieved precisely because of the regularization. Second, regularization greatly reduces the number of effective degrees of freedom, driving the eigenvalues that are small relative to λ essentially to zero. Empirically, a model with hundreds or thousands of observations – which could theoretically support as many degrees of freedom – often turns out to have on the order of 5-10 effective degrees of freedom. This ability to approximate complex functions, but with a preference for less complicated ones, is central to the wide applicability of KRLS. It makes models as complicated as needed, but not more so, and gains from the efficiency boost when simple models are sufficient. As we show below, the regularization can rescue so much efficiency that the resulting KRLS model is not much less efficient as a linear model even on linear data.

4.1.4. Finite Sample and Asymptotic Distribution of $\hat{y}$

Here, we establish asymptotic normality of the KRLS estimator. 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 (Normality in finite samples) *Under assumptions 1-5, $\hat{y} \sim N(y^\star, (\sigma_\epsilon K(K+\lambda I)^{-1})^2)$. The proof is given in the appendix.*

Second, we establish that the estimator is also normal asymptotically even when ϵ is non-normal but independently drawn from a distribution with finite mean and variance.

Assumption 6 (Regularity Conditions 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 (Asymptotic Normality) *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 given in the appendix. The resulting asymptotic distribution used for inference on any given $\hat{y}_i$ is*

$$\frac{\hat{y}_i - y_i^\star}{\sigma_\epsilon (K(K+\lambda I)^{-1})_{(i,i)}} \xrightarrow{d} N(0,1), \tag{13}$$

Theorem 4 is corroborated by simulations, which show that 95% confidence intervals based on standard errors computed by this method (a) closely match confidence intervals constructed from a non-parametric bootstrap and (b) have accurate empirical coverage rates under repeated sampling where new noise vectors are drawn for each iteration.

Taken together, these new results demonstrate the desirable theoretical properties of the KRLS estimator for the conditional expectation: it is unbiased for the best-fitting approximation to the true CEF in a large space of (penalized) functions (Theorems 1 and 2), consistent (Theorem 3), and asymptotically normally distributed given standard regularity conditions (Theorems 4 and 5). Moreover, variances can be estimated in closed form (Lemmas 1-3).

4.2. Interpretation and Quantities of Interest

One important benefit of KRLS over many other flexible modeling approaches is that the fitted KRLS model lends itself to a range of interpretational tools, which we develop in this section.

4.2.1. Estimating $E[y|X]$ and First Differences

The most straightforward interpretive element of KRLS is that we can use it to estimate the expectation of y conditional on $X = x$. From here, we can compute many quantities of interest such as first differences or marginal effects. We can also produce plots that show how the predicted outcomes change across a range of values for a given predictor variable while holding the other predictors fixed. For example, we can construct a dataset in which one predictor $x^{(a)}$ varies across a range of test values and the other predictors remain fixed at some constant value (e.g. the means) and then use this dataset to generate predicted outcomes, add a confidence envelope, and plot them against $x^{(a)}$ to explore ceteris paribus changes. Similar plots are typically used to interpret GAM models; however, the advantage of KRLS is that the learned model that is used to generate predicted outcomes does not rely on the additivity assumptions typically required for GAMs. Our companion software includes an option to produce such plots.

4.2.2. Partial Derivatives

We derive an estimator for the pointwise partial derivatives of y with respect to any particular input variable, $x^{(a)}$, which allows researchers to directly explore the pointwise marginal effects of each input variable and summarize them, for example, in the form of a regression table. Let $x^{(d)}$ be a particular variable, such that $X = [x^1 \dots x^d \dots x^D]$. Then, for a single observation, j, the partial derivative of y with respect to variable d is given by

$$\frac{\partial y}{\partial x_j^{(d)}} \approx \frac{-2}{\sigma^2} \sum_i c_i e^{\frac{-||x_i - x_j||^2}{\sigma^2}} (x_i^{(d)} - x_j^{(d)}). \tag{14}$$

The KRLS pointwise partial derivatives may vary across every point in the covariate space. One way to summarize the partial derivatives is to take their expectation. We, thus, construct the sample-average partial derivative of y with respect to $x^{(d)}$ at each observation:[14]

$$E_N \left[\frac{\partial y}{\partial x_j^{(d)}} \right] \approx \frac{-2}{\sigma^2 N} \sum_j \sum_i c_i e^{\frac{-||x_i - x_j||^2}{\sigma^2}} (x_i^{(d)} - x_j^{(d)}). \tag{15}$$

We also derive the variance of this quantity, and our software computes the pointwise and the sample-average partial derivative for each input variable together with their standard errors. The benefit of the sample-average partial derivative estimator is that it reports something akin to the usual β produced by linear regression: an estimate of the average marginal effect of each independent variable. However, there is a key difference between taking a best linear approximation to the data (as in OLS), versus fitting the CEF flexibly and then taking the average partial derivative in each dimension (as in KRLS). OLS gives a linear summary, but is highly susceptible to misspecification bias, in which the un-modeled effects of some observed variables can be mistakenly attributed to other observed variables. KRLS is much less susceptible to this bias, because it first fits the CEF more flexibly, and then can report back an average derivative over this improved fit.

Since KRLS provides partial derivatives for every observation, it allows for interpretation beyond the sample-average partial derivative. Plotting histograms of the pointwise derivatives, or plotting the derivative of y with respect to $x_i^{(d)}$ as a function of $x^{(d)}$, are useful interpretational tools. Plotting a histogram of $\frac{\partial y}{\partial x_i^{(d)}}$ over all i can quickly give the investigator a sense of whether the effect of a particular variable is relatively constant or very heterogeneous. It may turn out that the $\frac{\partial y}{\partial x^{(d)}}$ is bimodal, having a marginal effect that is strongly positive for one group of observations and strongly negative for another group. While the average partial derivative (or a β coefficient) would return a result near zero, this would obscure the fact that the variable in question is having a strong effect

[14] Such an average marginal effect has also been proposed by Long (1997) for use in GLMs, including those with interaction terms. This has the benefit of factoring the distribution of the data into the summary statistic.

but in opposite directions depending on the level of other variables. KRLS is well-suited to detect such effect heterogeneity. Our companion software includes an option to plot such histograms, as well as a range of other quantities.

4.2.3. Binary Independent Variables

KRLS works well with binary independent variables; however, they must be interpreted by a different approach than continuous variables. Given a binary variable $x^{(b)}$, the pointwise partial derivative $\frac{\partial y}{\partial x_i^{(b)}}$ is only observed where $x_j^{(b)} = 0$ or where $x_j^{(b)} = 1$. The partial derivatives at these two points do not characterize the expected effect of going from $x^{(b)} = 0$ to $x^{(b)} = 1$.[15] If the investigator wishes to know the expected difference in y between a case in which $x^{(b)} = 0$ and one in which $x^{(b)} = 1$, as is usually the case, we must instead compute first-differences directly. Let all other covariates (besides the binary covariate in question) be given by X. The first-difference sample estimator is $\frac{1}{N}\sum[\hat{y}_i|x_i^{(b)} = 1, X = x_i] - \frac{1}{N}\sum[\hat{y}_i|x_i^{(b)} = 0, X = x_i]$. This is computed by taking the mean $\hat{y}$ in one version of the dataset in which all X's retain their original value and all $x^{(b)} = 1$, and by subtracting from this the mean $\hat{y}$ in a dataset where all the values of $x^{(b)} = 0$. In the appendix, we derive closed-form estimators for the standard errors for this quantity. Our companion software detects binary variables and reports the first-difference estimate and its standard error, allowing users to interpret these effects as they are accustomed to from regression tables.

4.3. $E[y|x]$ returns to $E[y]$ for extreme examples of x

One important result is that KRLS protects against extrapolation for modeling extreme counterfactuals. Suppose we attempt to model a value of $\hat{y}_j$ for a test point x_j. If x_j lies far from all the observed data points, then $k(x_i, x_j)$ will be close to zero for all i. Thus, by equation (2), $f(x_j)$ will be close to zero, which also equals the mean of y due to pre-processing. Thus, if we attempt to predict $\hat{y}$ for a new counterfactual example that is far from the observed data, our estimate approaches the sample mean of the outcome variable. This property of the estimator is both useful and sensible. It is useful because it protects against highly model-dependent counterfactual reasoning based on extrapolation. In linear models, for example, counterfactuals are modeled as though

[15] The predicted function that KRLS fits for a binary input variable is a sigmoidal curve, less steep at the two endpoints than at the (unobserved) values in-between. Thus, the sample-average partial derivative on such variables will underestimate the marginal effect of going from 0 to 1 on this variable.

the linear trajectory of the conditional expectation function continues on indefinitely, creating a risk of producing highly implausible estimates (King and Zeng 2006). This property is also sensible, we argue, because in a Bayesian sense it reflects the knowledge that we have for extreme counterfactuals. Recall that under the similarity-based view, the only information we need about observations is how similar they are to other observations; the matrix of similarities, K, is a sufficient statistic for the data. If an observation is so unusual that it is not similar to any other observation, our best estimate of $E[\hat{y}_j|X = x_j]$ would simply be $E[y]$, as we have no basis for updating that expectation.

5. Simulation Results

Here, we show several simulation examples of KRLS that illustrate certain aspects of its behavior. Further examples are presented in the web appendix.

5.1. Leverage Points

One weakness of OLS is that a single aberrant data point can have an overwhelming effect on the coefficients and lead to unstable inferences. This concern is mitigated in KRLS, owing to the complexity-penalized objective function: adjusting the model to accommodate a single aberrant point typically adds more in complexity than it makes up for by improving model fit. To test this, we consider a linear data-generating process, $y = 2x + \epsilon$. In each simulation we draw $x \sim Unif(0, 1)$ and $\epsilon \sim N(0, .3)$. We then contaminate the data by setting a single data point to $(x = 5, y = -5)$, which is off the line described by the target function. As shown in the left panel of Figure 2, this single bad leverage point strongly biases the OLS estimates of the average marginal effect of x downwards (open circles), while the estimates of the average derivative from KRLS are robust even at small sample sizes (closed circles).

5.2. Efficiency Comparison

We expect that the added flexibility of KRLS would reduce the bias due to misspecification error, but come at the cost of increased variance, due to the usual bias-variance tradeoff. While true, regularization helps to prevent KRLS from suffering this problem too severely. The regularizer imposes a high penalty on complex, high-frequency functions, effectively reducing the space of functions and ensuring that small variations in the data do not lead to large variations in the fitted function. Thus, it reduces the variance. We illustrate this using a linear data-generating process, $y = 2x + \epsilon$, $x \sim N(0,1)$, and $\epsilon \sim N(0,.25)$, such that OLS is guaranteed to be the most efficient unbiased linear estimator by the Gauss-Markov theorem. The right panel in Figure 2 compares the standard error of the sample average partial derivative estimated by KRLS to that of $\hat{\beta}$ obtained by OLS. As expected, KRLS is not as efficient as OLS. However, the efficiency cost is quite modest, with the KRLS standard error on average being 13% larger than the standard errors from OLS. The efficiency cost is relatively low due to regularization, as discussed above. Both OLS and KRLS standard errors decrease at the rate of roughly $1/\sqrt{N}$ as suggested by our consistency result.

5.3. Over-fitting

A possible concern with flexible estimators is that they may be prone to overfitting, especially in large samples. With KRLS, regularization helps to prevent over-fitting by explicitly penalizing complex functions. To demonstrate this point, we consider a high-frequency function given by $y = .2\,sin(12\pi x) + sin(2\pi x)$ and run simulations with $x \sim Unif(0,1)$ and $\epsilon \sim N(0,.2)$ with two sample sizes $N = 40$ and $N = 400$. The results are displayed in the left panel of Figure 3. We find that for the small sample size, KRLS approximates the high-frequency target function (solid line) well with a smooth low-frequency approximation (dashed line). This approximation remains stable at the larger sample size (dotted line), indicating that KRLS is not prone to over-fit the function even as N grows large. This admittedly depends on the appropriate choice of λ, which is automatically chosen in all examples by GCV as described above.

5.4. Non-Smooth Functions

One potential downside of regularization is that KRLS is not well-suited to estimate discontinuous target functions. In the right panel of Figure 3, we use the same setup from the over-fitting simulation above, but replace the high-frequency function with a discontinuous step function. KRLS does not approximate the step well at $N = 40$, and the fit improves only modestly at $N = 400$, still failing to approximate the sharp discontinuity. However, KRLS still performs much better than the comparable OLS estimate, which uses x as a continuous regressor. The fact that KRLS tries to approximate the step with a smooth function is expected and desirable. For most social science problems, we would assume that the target function is continuous in the sense that very small changes in the independent variable are not associated with dramatic changes in the outcome variable, which is why KRLS uses such a smoothness prior by construction. Of course, if the discontinuity is known to the researcher, it should be directly incorporated into the KRLS or the OLS model by using a dummy variable $x' = \mathbf{1}[x > .5]$ instead of the continuous x regression. Both methods would then exactly fit the target function.

5.5. Interactions

We now turn to multivariate functions. First, we consider the standard interaction model where the target function is $y = .5 + x_1 + x_2 - 2(x_1 \cdot x_2) + \varepsilon$ with $x_j \sim Bernoulli(.5)$ for $j = 1, 2$ and $\varepsilon \sim N(0, .5)$. We fit KRLS and OLS models that include x_1 and x_2 as covariates and test the out-of-sample performance using the R^2 for predictions of $\hat{y}$ at $1,000$ test points that are drawn from the same distribution as the covariates. The upper panel in Figure 4 shows the out-of-sample R^2 estimates. KRLS (closed circles) accurately learns the interaction from the data and approaches the true R^2 as the sample size increases. OLS (open circles) misses the interaction and performs poorly even as the sample size increases.

Of course, in this simple case, we could get the correct answer with OLS if we specify the saturated regression that includes the interaction term $(x_1 \cdot x_2)$. However, even if the investigator suspects such an interaction needs to be modeled, the strategy of including interaction terms very quickly runs up against the combinatorial explosion of potential interactions in more realistic cases with multiple predictors. Consider a similar simulation for a more realistic case with ten binary predictors and a target function that contains several interactions: $y = (x_1 \cdot x_2) - 2(x_3 \cdot x_4) + 3(x_5 \cdot x_6 \cdot x_7) - (x_1 \cdot x_8) + 2(x_8 \cdot x_9 \cdot x_{10}) + x_{10}$. Here, it is difficult to search through the myriad of different OLS specifications to find the correct model: it would take 2^{10} terms to account for all the unique possible multiplicative interactions. This is why, in practice, social science researchers typically include no or very few interactions in their regressions. It is well-known that this results in often severe misspecification bias if the effects of some covariates depend on the levels of other covariates (e.g. Brambor et al. 2006). KRLS allows researchers to avoid this problem since it learns the interactions from the data.

The lower panel in Figure 4 shows that in this more complex example, the OLS regression that is linear in the predictors (open circles) performs very poorly, and this performance does not improve further as the sample size increases. Even at the largest sample size, it still misses close to half of the systematic variation in the outcome that results from the covariates. In stark contrast, the KRLS estimator (closed circles) performs well even at small sample sizes when there are fewer observations than the number of possible two-way interactions (not to mention higher-order interactions or higher-order powers of the input variables). Moreover, the out-of-sample performance approaches the true R^2 as the sample size increases, indicating that the learning of the function continues as the sample size grows larger. This clearly demonstrates how KRLS obviates the need for tedious specification searches and guards against misspecification bias. The KRLS estimator accurately learns the target function from the data and captures complex non-linearities or interactions that are likely to bias OLS estimates.

5.6. The Dangers of OLS with Multiplicative Interactions

Here, we show how the strategy of adding interaction terms can easily lead to incorrect inferences even in simple cases. Consider two correlated predictors $x_1 \sim Unif(0,2)$ and $x_2 = x_1 + \xi$ with $\xi \sim N(0,1)$. The true target function is $y = 5x_1^2$ and, thus, only depends on x_1 with a mild non-linearity. This non-linearity is so mild that in reasonable noisy samples, even a careful researcher that follows the textbook recommendations and first inspects a scatterplot between the outcome and x_1 might easily mistake it for a linear relationship. The same is true for the relationship between the outcome and the (conditionally irrelevant) predictor x_2. Given this, a researcher that has no additional knowledge about the true model is likely to fit a rather "flexible" regression model with a multiplicative interaction term given by $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3(x_1 \cdot x_2)$. To examine the performance of this model, we run a simulation that adds random noise and fits the model using outcomes generated by $y' = 5x_1^2 + \varepsilon$ where $\varepsilon \sim N(0,2)$.

The second column in Table 1 displays the coefficient estimates from the OLS regression (averaged across the simulations) together with their bootstrapped standard errors. In the eyes of the researcher, the OLS model performs rather well. Both lower-order terms and the interaction term are highly significant, and the model fit is good with $R^2 = .89$. In reality, however, using OLS with the added interaction term leads us to entirely false conclusions. We conclude that x_1 has a positive effect, and the magnitude of this effect increases with higher levels of x_2. Similarly, x_2 appears to have a negative effect at low levels of x_1 and a positive effect at high levels of x_1. Both conclusions are false and an artefact of misspecification bias. In truth no interaction effect exists – the effect of x_1 only depends on levels of x_1 – and x_2 has no effect at all.

The third column in Table 1 displays the estimates of the average pointwise derivatives from the KRLS estimator, which accurately recover the true average derivatives. The magnitude of the average marginal effect of x_2 is zero and highly insignificant. The average marginal effect of x_1 is highly significant and estimated at 9.2, which is fairly accurate given that x_1 is uniform between 0 and 2 (and, thus, we expect an average marginal effect of 10). Moreover, KRLS gives us more than just the average derivatives: it allows us to examine the effect heterogeneity by examining the marginal distribution of the pointwise derivatives. The next three columns display the 1st, 2nd, and 3rd quartile of the distributions of the marginal effects of the two predictors. The marginal effect of x_2 is close to zero throughout the support of x_2, which is accurate given that this predictor is indeed irrelevant for the outcome. The marginal effect of x_1 varies greatly in magnitude from about 5 at the 1st quartile to more than 14 at the 3rd quartile. This accurately captures the non-linearity in the true effect of x_1. (Figure A.4 in the appendix shows that the pointwise derivative estimates from KRLS accurately track the true marginal effects across the support of both predictors, while the marginal effects from the OLS interaction model produce misleading conclusions.)

5.7. Common Interactions and Non-additivity

Here, we show how KRLS is well-suited to fit target functions that are non-additive and/or involve more complex interactions as they arise in social science research. For the sake of presentation, we focus on target functions that involve two independent variables, but the principles generalize to higher-dimensional problems. We consider three types of functions: those with one "hill" and one "valley," two hills and two valleys, or three hills and three valleys (see appendix, Figures A.5, A.6 and A.6, respectively). These functions – especially the first two – correspond to rather common scenarios in the social sciences where the effect of one variable changes or dissipates depending on the effect of another. We simulate each type of function, using 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 simulations. In the appendix, we provide further explanation and visualizations pertaining to each simulation.

Table 2 displays both the in-sample and out-of-sample R^2 (based on 200 test points drawn from the same distribution as the training sample) for all three target functions and estimators. KRLS provides better in- and out-of-sample fits for all three target functions and the out-of-sample R^2 for each model is close to the true R^2 one would obtain knowing the functional form. These simulations increase our confidence that KRLS can capture complex non-linearity, non-additivity, and interactions that we may expect in social science data. While such features may be easy to detect in examples like these that only involve two predictors, they are even more likely in higher-dimensional problems where complex interactions and non-linearities are very hard to detect using plots or traditional diagnostics.

5.8. Comparison to Other Approaches

KRLS is not a panacea for all that ails empirical research, but our proposition is that it provides a useful addition to the empirical toolkit of social scientists, especially those currently using GLMs, owing to (a) the appropriateness of its assumptions to social science data, (b) its ease of use, and (c) the interpretability and ease with which relevant quantities of interest and their variances are produced. It therefore fulfills different needs than many other machine learning or flexible modeling approaches, such as neural networks, regression trees, k-Nearest Neighbors, SVMs, and GAMs, to name a few. In the appendix, we describe in greater detail how KRLS compares to important classes of models on interpretability and inference, with special attention to Generalized Additive Models (GAMs), and to approaches that involve explicit basis expansions followed by fitting methods that force many of the coefficients to be exactly zero (LASSO). At bottom, we do not claim that KRLS is generally superior to other approaches, but rather that it provides a particularly useful marriage of flexibility and interpretability. It does so with far lower risk of misspecifciation bias then highly constrained models, while minimizing arbitrary choices about basis expansions and selection of smoothing parameters.

These differences aside, in proposing a new method it is useful to compare its pure modeling performance to other candidates. In this area, KRLS does very well.[16] To further illustrate how KRLS compares against other methods that have appeared in political science, we replicate a simulation from Wood (2003) that was designed specifically to illustrate the use of GAMs. The data-

[16] The RLS models on which KRLS is based have proven effective even when used for classification rather than regression, with performance indistinguishable from state-of-the-art Support Vector Machines (Rifkin et al. 2003).

generating process is given by $x_1, x_2 \sim Unif(0,1)$, $\epsilon \sim N(0,.25)$, and $y = e^{10(-(x_1-.25)^2-(x_2-.25)^2)} + .5 * e^{14(-(x_1-.7)^2-(x_2-.7)^2)} + \epsilon$. We consider five models: (1) KRLS with default choices ($\sigma^2 = D = 2$), implemented in our R package simply as `krls(y=y,X=cbind(x1,x2))`; (2) a "naive" GAM (GAM1) that smoothes x_1 and x_2 separately, but then assumes that they add; (3) a "smart" GAM (GAM2) that smoothes x_1 and x_2 together using the default thin-plate splines and the default method for choosing the number of basis functions in the `mgcv` package in R; (4) a generous linear model (LM), $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 \times x_2$; and (5) a neural network (NN) with 5 hidden units and all other parameters at their defaults using the `NeuralNet` package in R. We train this model on samples of 50, 100, or 200 observations, and then test it on 100 out-of-sample observations. Results for the RMSE of each model averaged over 200 iterations at each sample size are shown in Table 3. KRLS performs as well as or better than all other methods, at all sample sizes. In smaller samples, it clearly dominates. As the sample size increases, the fully smoothed GAM performs very similarly.[17]

6. Application: Predicting Genocide

In this section we show an application of KRLS to a real data example. A second empirical example that draws on Brambor et al. (2006) is provided in the web appendix.

In a widely cited article, Harff (2003) examines data from 126 political instability events (i.e. internal wars and regime changes away from democracy) to determine which factors can be used to predict whether a state will commit genocide.[18] Harff proposes a "structural model of genocide" where a dummy for genocide onset is regressed on two continuous variables, *prior upheaval* (summed years of prior instability events in the past 15 years) and *trade openness* (imports and exports as a fraction of GDP in logs), and four dummy variables that capture if the state is an *autocracy*, had a *prior genocide*, and whether the ruling elite has an *ideological character*, and/or an *ethnic character*.[19] The first column in Table 4 replicates the original specification using a linear probability

[17] KRLS and GAMs in which all variables are smoothed together are fundamentally similar. The main difference under current implementations (our package for KRLS and `mgcv` for GAMs) include the following: (1) the fewer interpretable quantities produced by GAMs; (2) the inability of GAMs to fully smooth together more than a few input variables; and (3) the kernel implied by GAMs that leads to straight-line extrapolation outside the support of X. These are discussed further in the web appendix.

[18] The American Political Science Association lists this paper as the 15th most downloaded paper in the *American Political Science Review.* According to Google Scholar, this article has been cited 310 times.

[19] See Harff (2003) for details. Notice that Harff dichotomized a number of continuous variables (such as the polity

model.[20] The next four columns on the left present the replication results from the KRLS estimator. We report first differences for all the binary input variables as described above. The analysis yields several lessons.

First, the in-sample R^2 from the original logit model and KRLS are very similar (32% versus 34%), but KRLS dominates in terms of its ROC curve for predicting genocide, with statistically significantly more area under the curve ($p < 0.03$). This superior performance (at least in-sample) is notable given that, as Harff reports, her final specification was selected after an extensive search through a large number of predictors and specifications. Here, KRLS explains more of the variation without any human specification search. The researcher simply passes the predictor matrix to KRLS, which learns the functional form from the data, and this improves empirical performance and reduces arbitrariness in selecting a particular specification.

Second, the average marginal effects reported by KRLS (shown in the second column) are all of reasonable size and tend to be in the same direction as but somewhat smaller than the estimates from the linear probability model. We also see some important differences. The OLS model shows a significant effect of *prior upheaval*, with a one standard deviation increase corresponding to a 10 percentage point increase in the probability of genocide onset. This completely vanishes in the KRLS model, which yields an average marginal effect of zero that is also highly insignificant. This sharply different finding is confirmed when we look beyond the average marginal effect. Recall that the average marginal effects, while a useful summary tool especially to compare to GLMs, are only summaries and can hide interesting heterogeneity in the actual marginal effects across the covariate space. To examine the effect heterogeneity, the next three columns on the left show the quartiles of the distribution of pointwise marginal effects for each input variable. Figure 5 also plots histograms to visualize the distributions. We see that the effect of *prior upheaval* is essentially zero at every point.

What explains this difference in the marginal effect estimate? It turns out that the significant effect in the OLS model is an artefact of misspecification bias. The variable *prior upheaval* is strongly right-skewed and when logged to make it more appropriate for linear regression the "effect"

score), which discards valuable information. With KRLS, one could instead use the original continuous variables unless there was a strong reason to code dummies. In fact, tests confirm that using the original continuous variables with KRLS results in a more predictive model.

[20] Harff used a logit model, but the linear probability model produces similar results.

disappears entirely. This showcases the general problem that misspecification bias is often difficult to avoid in typical political science data, even for experienced researchers that publish in top journals and engage in various model diagnostics and specification searches. It also highlights the advantages of a more flexible approach such as KRLS, which avoids misspecification bias while yielding marginal effects estimates that are as easy to interpret as OLS and also make richer interpretation possible.

Third, further exploring the effect heterogeneity, we find that for some variables, such as *priorgen* and *autocracy*, the marginal effect lies to the same side of zero at almost every point, indicating that these variables have similar marginal effects in the same direction regardless of their level or the levels of other variables. We also find interesting results for *ethnic character* and *ideological character*. The histograms of their marginal effects (measured as first-differences) suggest that these variables have highly heterogeneous effects that depend on the levels of other variables in the model. By regressing the marginal effects for these variables on all the original variables, we can quickly notice a number of likely interactions.[21] Specifically, the pointwise marginal effects of *ethnic character* and *ideological character* are strongly negatively correlated with the level of *trade openness*. Splitting the sample at the median value of *trade openness*, we find that *ideological character* is associated with an increase of 9 percentage points in the probability of genocide onset when *trade openness* is low, but only a 3 percentage point increase when *trade openness* is high. Similarly, *ethnic character* is associated with an increase of 5 percentage points among the low *trade openness* observations, but has no marginal effect among the high *trade openness* observations. These findings are of substantive interest: for states with higher trade openness, the relationship between ethnic or ideological ruling elites and genocide is not as strong as it is in states with low trade openness. Since KRLS also provides closed-form estimates of the variance-covariance matrix for each of the pointwise partial derivatives, researchers could go further and test hypotheses or construct standard errors if desired.

This brief example demonstrates that KRLS is appropriate and effective in dealing with real-world data even in relatively small datasets. While KRLS offers much more flexibility than GLMs

[21] This approach is helpful to identify non-linearities and interaction effects. For each variable, take the pointwise partial derivatives (or first-differences) modeled by KRLS and regress them on all original independent variables to see which of them help explain the marginal effects. For example, if $\frac{\partial y}{\partial x^{(a)}}$ is found to be well-explained by $x^{(a)}$ itself, then this suggests a non-linearity in $x^{(a)}$ (because the derivative changes with the level of the same variable). Likewise, if $\frac{\partial y}{\partial x^{(a)}}$ is well-explained by another variable $x^{(b)}$, this suggests an interaction effect (the marginal effect of one variable, $x^{(a)}$ depends on the level of another, $x^{(b)}$).

and guards against misspecification bias, it is straightforward to interpret the results in ways that are familiar to researchers from GLMs.

7. Conclusion

To date, it has been difficult to find user-friendly approaches that avoid the dangers of misspecification while also generating quantities of interest that are as interpretable and appealing as the coefficients from linear models. We argue that KRLS represents a particularly useful marriage of flexibility and interpretability, especially for current GLM users looking for more powerful modeling approaches. It allows users to model non-linear and non-additive effects, minimizing misspecification bias, whilst still producing quantities of interest that allow the investigator to make "simple" interpretations (similar to those allowed by GLMs) and, if desired, to take a closer look at non-constant marginal effects. These quantities are produced by a single run of the model, and the model does not require user input regarding functional form or parameter settings, improving falsifiability. We have illustrated how KRLS accomplishes this improved tradeoff between flexibility and interpretability by starting from a different set of assumptions altogether: rather than assume that the target function is well-fitted by a linear combination of the original regressors, it is instead modeled in an N-dimensional space using information about similarity to each observation, but with a preference for less complicated functions, improving stability and efficiency. Since KRLS uses global information from all data points, it is less susceptible to the curse of dimensionality than purely local methods such as k-nearest neighbors and matching.

We have established a number of desirable properties of this technique. First, it allows computationaly tractable, closed-form solutions for many quantities including $E[y|X]$, the variance of this estimator, the pointwise partial derivatives with respect to each variable, the sample average partial derivatives, and their variances. We have also shown that it is unbiased, consistent, and asymptotical normal. Simulations have demonstrated the performance of this method, even in small samples and high-dimensional spaces. They have also shown that even when the true data-generating process is linear, the KRLS estimate of the average partial derivative is not much less efficient than the analogous OLS coefficient, and far more robust to bad leverage points.

We believe KRLS is broadly useful whenever investigators are unsure of the functional form in regression and classification problems. However, there remains considerable room for further research. Our hope is that the approach provided here and our companion software will allow more researchers to begin using KRLS or methods like it; only when when tested by a larger community of scholars will we be able to determine the true usefulness of this approach. Specific research tasks remain as well. Due to the memory demands of forming and working with an $N \times N$ matrix, the practical limit on N for most users is currently in the tens of thousands. Work on the best ways of resolving this constraint would be useful. In addition, the most effective methods for choosing λ and σ^2 are still relatively open questions, and it would be useful to develop heteroscedasticity-, autocorrelation-, and cluster-robust estimators for standard errors.

References

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.

De Vito, E., Caponnetto, A. and Rosasco, L. (2005). Model selection for regularized least-squares algorithm in learning theory, *Foundations of Computational Mathematics* **5**(1): 59–85.

Evgeniou, T., Pontil, M. and Poggio, T. (2000). Regularization networks and support vector machines, *Advances in Computational Mathematics* **13**(1): 1–50.

Friedrich, R. J. (1982). In defense of multiplicative terms in multiple regression equations, *American Journal of Political Science* pp. 797–833.

Golub, G. H., Heath, M. and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter, *Technometrics* **21**(2): 215–223.

Harff, B. (2003). No lessons learned from the holocaust? assessing risks of genocide and political mass murder since 1955, *American Political Science Review* **97**(1): 57–73.

Jackson, J. E. (1991). Estimation of models with variable coefficients, *Political Analysis* **3**(1): 27–49.

Kimeldorf, G. and Wahba, G. (1970). A correspondence between bayesian estimation on stochastic processes and smoothing by splines, *The Annals of Mathematical Statistics* **41**(2): 495–502.

King, G. and Zeng, L. (2006). The Dangers of Extreme Counterfactuals, *Political Analysis* **14**(2).

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*, Vol. 7, Sage Publications, Incorporated.

Rifkin, R. M. and Lippert, R. A. (2007). Notes on regularized least squares, *Technical report*, MIT Computer Science and Artificial Intelligence Laboratory Technical Report.

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.

Saunders, C., Gammerman, A. and Vovk, V. (1998). Ridge regression learning algorithm in dual variables, *Proceedings of the 15th International Conference on Machine Learning*, Vol. 19980, San Frsncisco, CA, USA: Morgan Kaufmann, pp. 515–521.

Schölkopf, B. and Smola, A. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*, the MIT Press.

Tychonoff, A. N. (1963). Solution of incorrectly formulated problems and the regularization method, *Doklady Akademii Nauk SSSR* **151**: 501504. Translated in Soviet Mathematics 4: 10351038.

Wood, S. N. (2003). Thin plate regression splines, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65**(1): 95–114.

Tables

Table 1: Comparing KRLS to OLS with Multiplicative Interactions

Estimator	OLS	KRLS			
$\partial y/\partial x_{ij}$	Average	Average	1st Qu.	Median	3rd Qu.
const	-1.50				
	(0.34)				
x_1	7.51	9.22	5.22	9.38	14.03
	(0.40)	(0.52)	(0.82)	(0.85)	(0.79)
x_2	-1.28	0.02	-0.08	0.00	0.10
	(0.21)	(0.13)	(0.19)	(0.16)	(0.20)
$(x_1 \times x_2)$	1.24				
	(0.15)				
N	250				

Note: Point estimates of marginal effects from OLS and KRLS regression with bootstrapped standard errors in parenthesis. For KRLS, the table shows the average and the quartiles of the distribution of the pointwise marginal effects. True target function is $y = 5x_1^2$ and simulated using $y' = 5x_1^2 + \varepsilon$ with $\varepsilon \sim (0, 2)$, $x_1 \sim Unif(0, 2)$, and $x_2 = x_1 + \xi$ with $\xi \sim N(0, 1)$. With OLS, we conclude that x_1 has a positive effect that grows with higher levels of x_2 and that x_2 has a negative (positive) effect at low (high) levels of x_1. The true marginal effects are $\frac{\partial y}{\partial x_1} = 10x_1$ and $\frac{\partial y}{\partial x_2} = 0$; the effect of x_1 only depends on levels of x_1 and x_2 has no effect at all. The KRLS estimator accurately recovers the true average derivatives. The marginal effects of x_2 are close to zero throughout the support of x_2. The marginal effects of x_1 varies from about 5 at the 1st quartile to about 14 at the 3rd quartile.

Table 2: KRLS Captures Complex Interactions and Non-additivity

Target Function	One Hill One Valley	Two Hills Two Valleys	Three Hills Three Valleys
In Sample R^2			
KRLS	0.75	0.41	0.52
OLS	0.61	0.01	0.01
GAM	0.63	0.21	0.05
Out-of-sample R^2			
KRLS	0.70	0.35	0.45
OLS	0.60	-0.01	-0.01
GAM	0.60	0.13	-0.03
True R^2	0.73	0.39	0.51

Note: In- and out-of-sample R^2 (based on 200 test points) for simulations using the three target functions displayed in Figures A.5, A.6, and A.7 inthe appendix with the OLS, GAM, and KRLS estimators. KRLS attains the best in-sample and out-of-sample fit for all three functions.

Table 3: Comparing KRLS to Other Methods

	Mean RMSE		
Model	N=50	N=100	N=200
KRLS	0.139	0.107	0.088
GAM2	0.143	0.109	0.088
NN	0.312	0.177	0.118
LM	0.193	0.177	0.169
GAM1	0.234	0.213	0.202

Note: Simulation comparing RMSE for out-of-sample fits generated by five models, averaged over 200 iterations. The data-generating process is based on Wood (2003): $x_1, x_2 \sim Unif(0,1)$, $\epsilon \sim N(0,.25)$, and $y = e^{10(-(x_1-.25)^2-(x_2-.25)^2)} + .5 * e^{14(-(x_1-.7)^2-(x_2-.7)^2)} + \epsilon$. Models are KRLS with default choices; a "naive" GAM (GAM1) that smoothes x_1 and x_2 separately; (3) a "smart" GAM (GAM2) that smoothes x_1 and x_2 together; (4) a generous linear model (LM), $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 \times x_2$; and (5) a neural network (NN) with 5 hidden units. Models are trained on samples of 50, 100, or 200 observations, and then tested on 100 out-of-sample observations. KRLS out-performs all other methods in small samples. In larger samples, KRLS and the GAM2 (with "full-smoothing") perform similarly. The linear model, despite including terms for x_1^2, x_2^2, and $x_1 x_2$, does not perform particularly well. GAM1 also performs poorly in all circumstances.

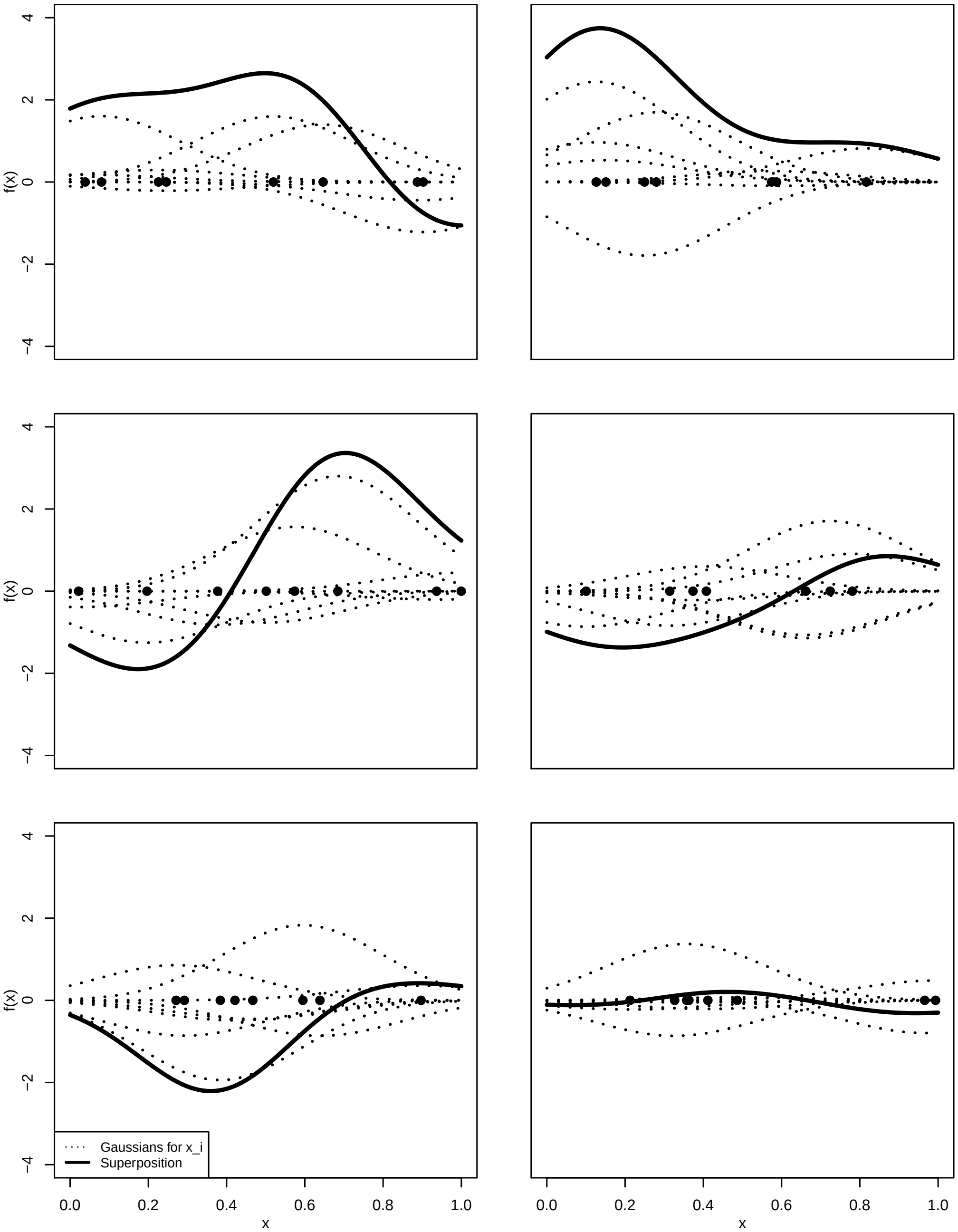
Table 4: Predictors of Genocide Onset: OLS versus KRLS

Estimator	OLS	KRLS			
		$\partial y / \partial x_{ij}$			
	β	Average	1st Qu.	Median	3rd Qu.
Prior upheaval	0.01*	0.00	-0.00	0.00	0.00
	(0.00)	(0.00)			
Prior genocide	0.26*	0.19*	0.14	0.23	0.27
	(0.12)	(0.08)			
Ideological char. of elite	0.15	0.12	0.09	0.14	0.19
	(0.08)	(0.08)			
Autocracy	0.16*	0.12	0.09	0.11	0.14
	(0.08)	(0.07)			
Ethnic char. of elite	0.12	0.05	0.01	0.05	0.08
	(0.08)	(0.08)			
Trade openness (log)	-0.17*	-0.09*	-0.14	-0.07	-0.05
	(0.06)	(0.03)			
Intercept	0.66				
	(0.22)				

Note: Replication of the "structural model of genocide" by Harff (2003). Marginal effects of predictors from OLS regression and KRLS regression with standard errors in parenthesis. For KRLS, the table shows the average of the pointwise derivative as well as the quartiles of their distribution to examine the effect heterogeneity. The dependent variable is a binary indicator for genocide onsets. N=126. *p-value $<$ .05. See text for details.

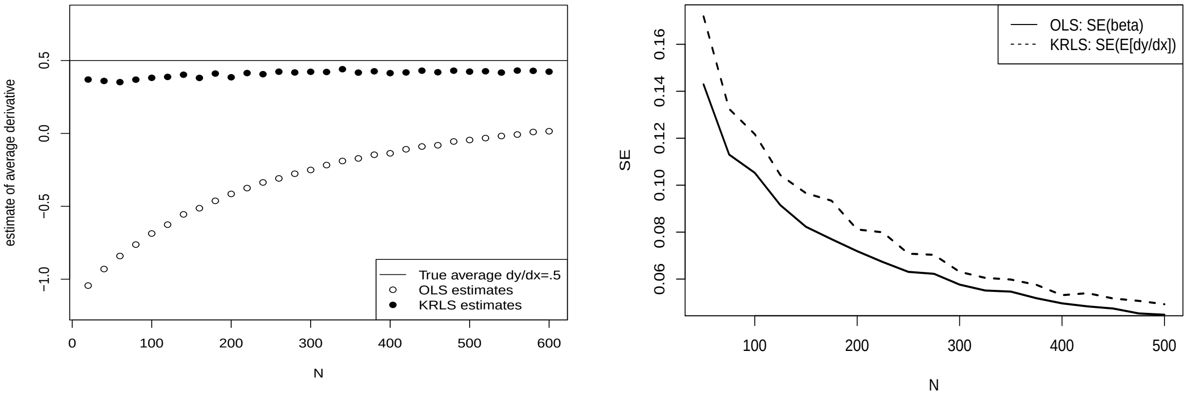
Figures

Figure 1: Random Samples of Functions of the Form $f(x) = \sum_{i=1}^{N} c_i k(x, x_i)$



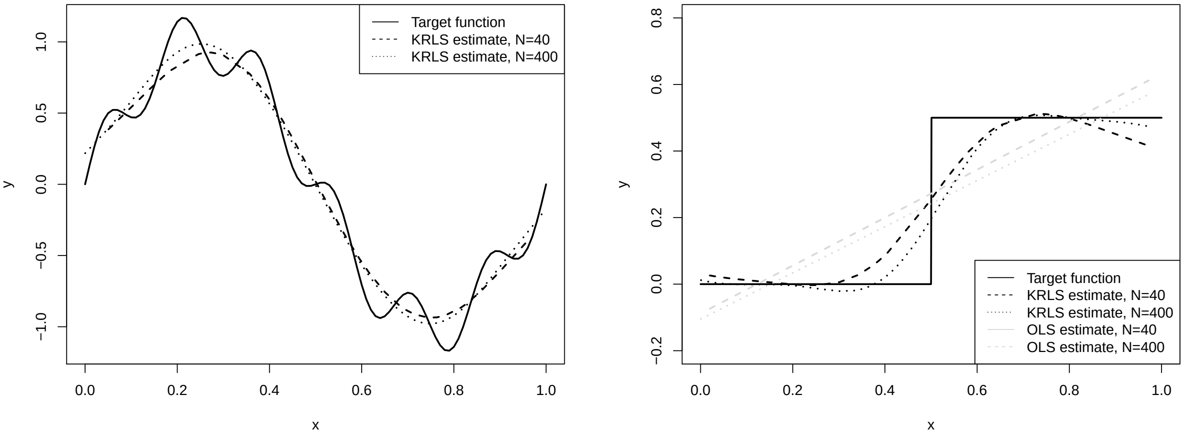
Note: The target function is created by centering a Gaussian over each x_i, scaling each by its c_i, and then summing them. We use 8 observations with $c_i \sim N(0,1)$, $x \sim Unif(0,1)$, and a fixed value for the bandwidth of the kernel σ^2. The dots represent the sampled data points, the dotted lines refer to the scaled Gaussian kernels that are placed over each sample point, and the solid lines represent the target functions created from the superpositions. Notice that the center of the Gaussian curves depends on the point x_i, its upwards or downward direction depends on the sign of the weight c_i, and its amplitude depends on the magnitude of the weight c_i (as well as the fixed σ^2).

Figure 2: KRLS Compares Well to OLS with Linear Data-generating Processes



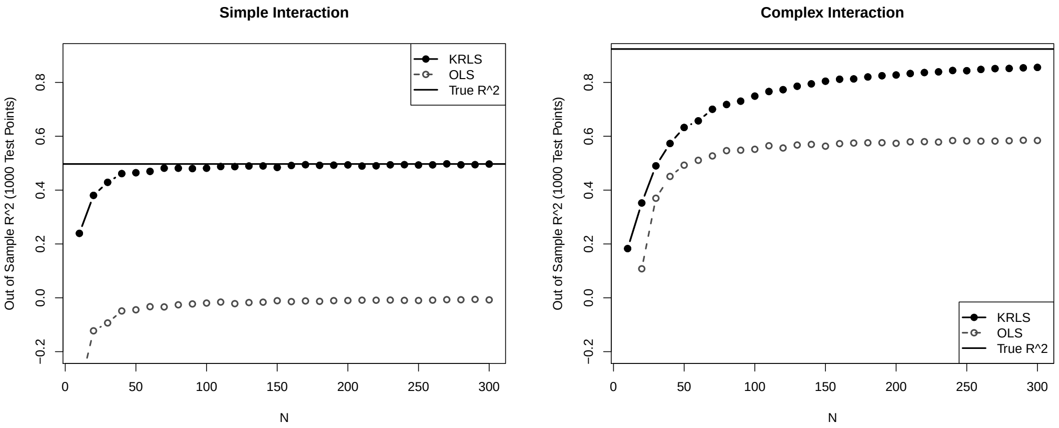
Left: Simulation to recover the average derivative of $y = .5x$, i.e. $\frac{\partial y}{\partial x} = .5$ (solid line). For each sample size, we run 100 simulations with observed outcomes $y = .5x + \varepsilon$ where $x \sim Unif(0,1)$ and $\varepsilon \sim N(0,.3)$. One contaminated data point is set to $(y_i = -5, x_i = 5)$. Dots represent the mean estimated average derivative for each sample size for OLS (open circles) and KRLS (full circles). The simulation shows that KRLS is robust to the bad leverage point, while OLS is not. *Right*: Comparison of standard error of β from OLS (solid line) to the standard error of the sample average partial derivative from KRLS (dashed line). Data are generated according to $y = 2x + \epsilon$, with $x \sim N(0,1), \epsilon \sim N(0,1)$ with 10 iterations for each sample size. KRLS is nearly as efficient as OLS at all but very small sample sizes, with standard errors on average approximately 13% larger than those of OLS.

Figure 3: KRLS With High-Frequency and Discontinuous Functions



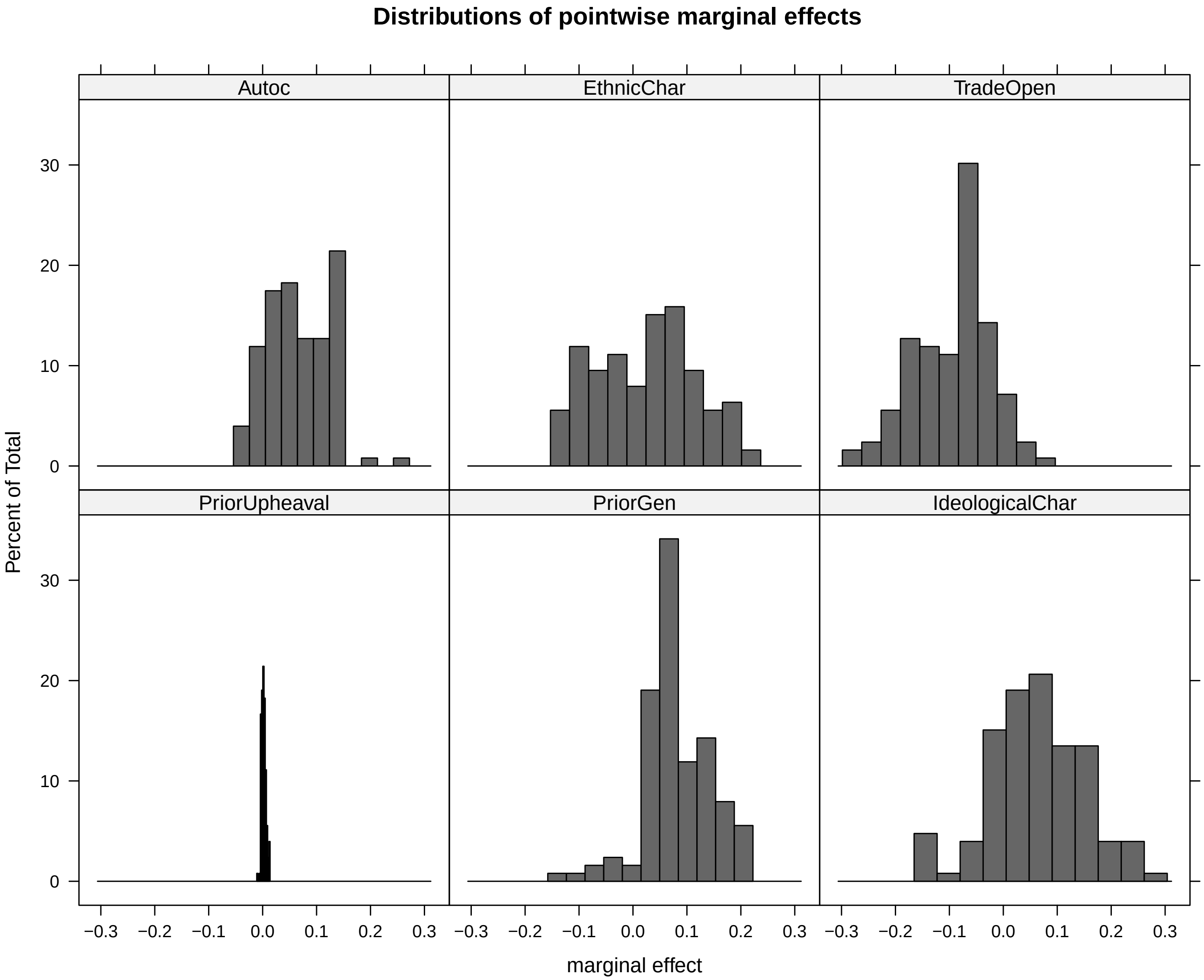
Left: Simulation to recover a high-frequency target function given by $y = .2 * sin(12\pi x) + sin(2\pi x)$ (solid line). For each sample size, we run 100 simulations where we draw $x \sim Unif(0,1)$ and simulate observed outcomes as $y = .2 * sin(12\pi x) + sin(2\pi x) + \varepsilon$ where $\varepsilon \sim N(0,.2)$. The dashed line shows mean estimates across simulations for N=40 and dotted line for N=400. Results show that KRLS finds a low-frequency approximation even at the larger sample sizes. *Right*: Simulation to recover discontinuous target function given by $y = .5 * \mathbf{1}(x > .5)$ (solid line). For each sample size, we run 100 simulations where we draw $x \sim Unif(0,1)$ and simulate observed outcomes as $y = .5 * \mathbf{1}(x > .5) + \varepsilon$ where $\varepsilon \sim N(0,.2)$. Dashed lines show mean estimates across simulations for N=40 and dotted lines for N=400. Results show that KRLS fails to approximate the sharp discontinuity even at the larger sample size, but still dominates the comparable OLS estimate, which uses x as a continuous regressor.

Figure 4: KRLS Learns Interactions from the Data



Simulations to recover target functions that include multiplicative interaction terms. *Left*: The target function is $y = .5 + x_1 + x_2 - 2(x_1 \cdot x_2) + \varepsilon$ with $x_j \sim Bernoulli(.5)$ for $j = 1, 2$ and $\varepsilon \sim N(0, .5)$. *Right*: Target function is $y = (x_1 \cdot x_2) - 2(x_3 \cdot x_4) + 3(x_5 \cdot x_6 \cdot x_7) - (x_1 \cdot x_8) + 2(x_8 \cdot x_9 \cdot x_{10}) + x_{10}$ where all x are drawn i.i.d. $Bernoulli(p)$ with $p = .25$ for x_1 and x_2, $p = .75$ for x_3 and x_4, and $p = .5$ for all others. For each sample size, we run 100 simulations where we draw the x and simulate outcomes using $y = y_{true} + \varepsilon$ where $\varepsilon \sim N(0, .5)$ for the training data. We use $1,000$ test points drawn from the same distribution to test the out-of-sample R^2 of the estimators. The closed circles show the average R^2 estimates across simulations for the KRLS estimator, the open circles show the estimates for the OLS regression that uses all x as predictors. The true R^2 is given by the solid line. The results show that KRLS learns the interactions from the data and approaches the true R^2 one would obtain knowing the functional form as the sample size increases.

Figure 5: Effect Heterogeneity in Harff Data



Histograms of pointwise marginal effects based on KRLS fit to Harff data (Model 2 in Table 4).