Advance Access publication October 10, 2013

Political Analysis (2014) 22:143–168
doi:10.1093/pan/mpt019

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

Jens Hainmueller
Department of Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139
e-mail: jhainm@mit.edu (corresponding author)

Chad Hazlett
Department of Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139
e-mail: hazlett@mit.edu

Edited by R. Michael Alvarez

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 nonlinearities, interactions, 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 a simple automated rule for choosing the kernel bandwidth, and (4) providing companion software. We illustrate the use of the method through simulations and empirical examples.

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 problematic in social science data, which are frequently ridden with nonlinearities, nonadditivity, 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, incorrect inferences, 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

Authors' note: The authors are listed in alphabetical order and contributed equally. We thank Jeremy Ferwerda, Dominik Hangartner, Danny Hidalgo, Gary King, Lorenzo Rosasco, Marc Ratkovic, Teppei Yamamoto, our anonymous reviewers, the editors, 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 article in `R`, `Matlab`, and `Stata` can be downloaded from the authors' Web pages. Replication materials are available in the *Political Analysis* Dataverse at http://dvn.iq.harvard.edu/dvn/dv/pan. The usual disclaimer applies. Supplementary materials for this article are available on the *Political Analysis* Web site.

(e.g., Friedrich 1982; Jackson 1991; Brambor, Clark, and Golder 2006). However, higher-order terms only allow for interactions of a prescribed type, and even for experienced researchers, it is typically very difficult to find the correct functional form 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. Although some more flexible methods, such as neural networks (NNs) (e.g., Beck, King, and Zeng 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 nontrivial modifications and often impracticable computational demands.

In this article, we describe Kernel Regularized Least Squares (KRLS). This approach draws from Regularized Least Squares (RLS), a well-established method in the machine learning literature (see, e.g., Rifkin, Yeo, and Poggio 2003).[1] We add the "K" to (1) emphasize that it employs kernels (whereas the term RLS can also apply to nonkernelized models), and (2) designate the specific set of choices we have made in this version of RLS, including procedures we developed to remove all parameter selection from the investigator's hands and, most importantly, methodological innovations we have added relating to interpretability and inference.

The 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 often less interpretable machine learning approaches. KRLS is an easy-to-use approach that 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. This method belongs to a class of models for which marginal effects are well-behaved and easily obtainable due to the existence of a continuously differentiable solution surface, estimated in closed form. It also readily admits to statistical inference using closed form expressions, and has desirable statistical properties under relatively weak assumptions. The resulting model is 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 to obtain an average partial derivative similar to a $\hat{\beta}$ coefficient from linear regression.

Because it marries flexibility with interpretability, the KRLS approach is suitable for a wide range of regression and classification problems where the correct functional form is unknown. This includes exploratory analysis to learn about the data-generating process, model-based causal inference, or prediction problems that require an accurate approximation of a 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 from the curse of dimensionality, and producing principled variance estimates in closed form. Finally,

[1] Similar methods appear under various names, including Regularization Networks (e.g., Evgeniou, Pontil, and Poggio 2000) and Kernel Ridge Regression (e.g., Saunders, Gammerman, and Vovk 1998).

although necessarily somewhat less efficient than Ordinary Least Squares (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 article are three-fold. 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 methodological innovations. We (1) derive closed-form estimators for pointwise and average marginal effects, (2) derive closed-form variance estimators for these quantities to enable hypothesis tests and the construction of confidence intervals, (3) establish the unbiasedness, consistency, and asymptotic normality of the estimator for fitted values under conditions more general than those required for GLMs, and (4) derive 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

RLS approaches with kernels, of which KRLS is a special case, 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 method works and why it is a good fit for many social science problems. Further below we also provide a more rigorous, but perhaps less intuitive, justification.[2]

2.1 *Similarity-Based View*

Assume that we draw i.i.d. data of the form (y_i, x_i), where $i = 1, \ldots, N$ 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 $k(\cdot, \cdot)$ that takes two arguments and produces a real-valued output.[3] It is useful to think of the kernel function as providing a measure of similarity between two input patterns. Although many kernels are available, the kernel used in KRLS and throughout this article is the Gaussian kernel given by

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

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 with σ^2 in place of $2\sigma^2$, and omitting the normalizing factor $1/\sqrt{2\pi\sigma^2}$. The most important feature of this kernel is that it reaches its maximum of one only when $x_i = x_j$ and grows closer to zero as x_i and x_j become more distant. 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]

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

[2]Another justification is based on the analysis of reproducing kernels, and the corresponding spaces of functions they generate along with norms over those spaces. For details on this approach, we direct readers to recent reviews included in Evgeniou, Pontil, and Poggio (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 nonparametric methods commonly called "kernel regression" that involve using a kernel to construct a weighted local estimate.

[4]Below we provide a formal justification for this space based on ridge regressions in high-dimensional feature spaces.

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

$$f(x^\star) = c_1 k(x^\star, x_1) + c_2 k(x^\star, x_2) + \ldots + c_N k(x^\star, x_N) \quad (3)$$

$$= c_1(\text{similarity of } x^\star \text{ to } x_1) + c_2(\text{sim. of } x^\star \text{ to } x_2) + \ldots + c_N(\text{sim. of } x^\star \text{ to } x_N). \quad (4)$$

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 information is 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 expect that the nearness of a given observation, x_i, to other observations reveals 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

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

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 eight data points $x_i \sim \text{Uniform}(0, 1)$ 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). This figure shows that the function space is much more flexible than the function spaces available to GLMs; it enables us to approximate highly nonlinear and nonadditive 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 data set, 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 refer to the c_i weights as *choice coefficients*, similar to the role that β coefficients play in linear regression. Notice that a great many choices of c_i can produce highly similar fits—a problem resolved in the next section through regularization. (In the supplementary appendix, we present a toy example to build intuition for the mechanics of fitting the function; see Fig. A1.)

Before describing how KRLS chooses the choice coefficients, we introduce a more convenient matrix notation. Let K be the $N \times N$ symmetric *Kernel matrix* whose jth, ith entry is $k(x_j, x_i)$; it measures the pairwise similarities between each of the N input patterns x_i. Let $c = [c_1, \ldots, c_N]^T$ be the $N \times 1$ vector of choice coefficients and $y = [y_1, \ldots, y_N]^T$ be the $N \times 1$ vector of outcome values.

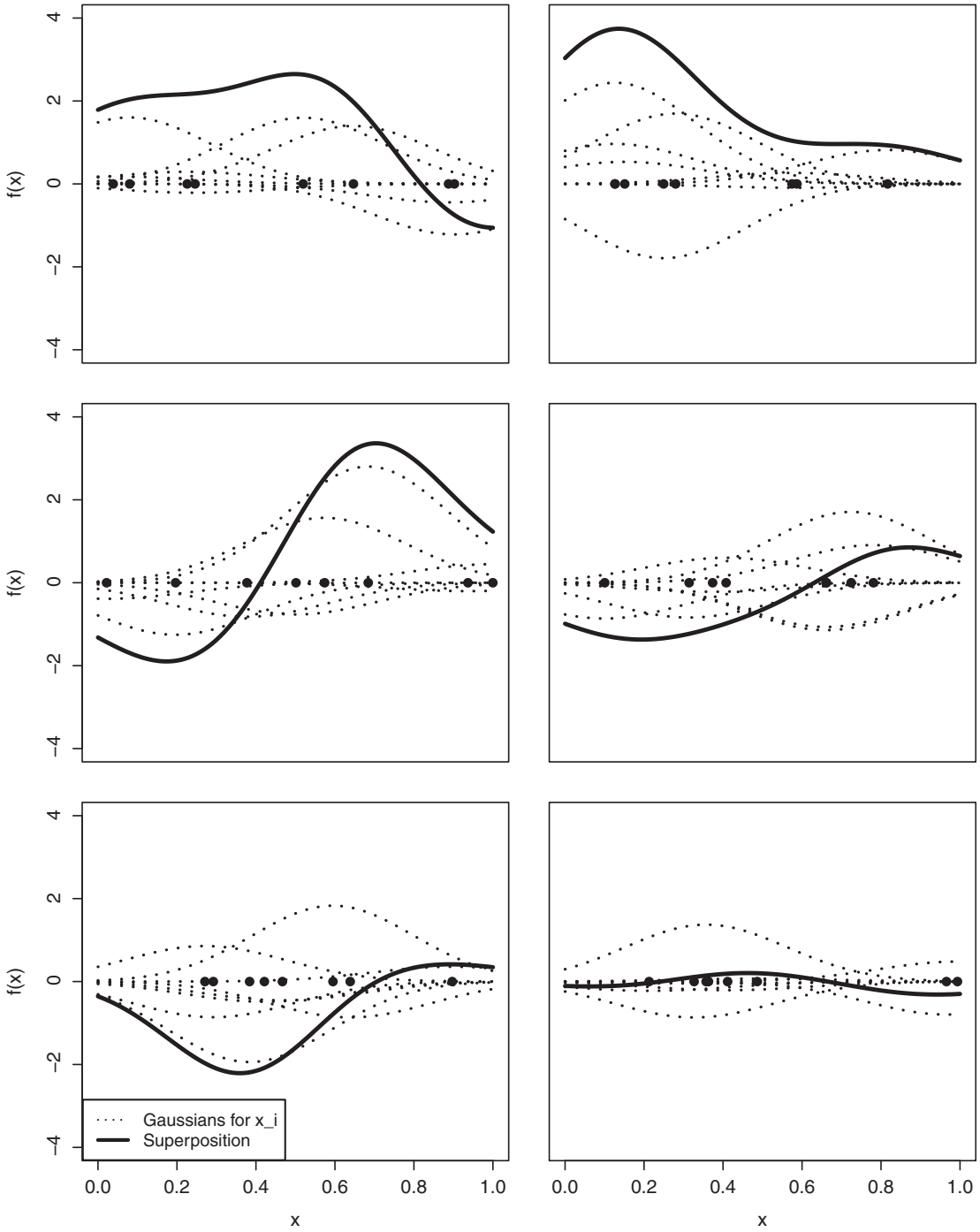

Fig. 1 Random samples of functions of the form $f(x) = \sum_{i=1}^{N} c_i k(x, x_i)$. The target function is created by centering a Gaussian over each x_i, scaling each by its c_i, and then summing them. We use eight observations with $c_i \sim N(0, 1)$, $x \sim \text{Unif}(0, 1)$, and a fixed value 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 upward 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).

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 (LM): 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 data set. Notice that the matrix K will be symmetric and positive semi-definite and, thus, invertible.[5] Therefore, there is a "perfect" solution to the linear system $y = Kc$, or equivalently, there is 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*

Although 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 underutilized 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 between the observed data points. 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 and useful than "high-frequency" relationships. (Figure A2 in the supplementary appendix provides an example for a low- and high-frequency explanation of the relationship between x and y.)[6]

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 trade-off 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 a higher

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

[6] 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 (e.g., a sigmoidal curve in the case of normally distributed noise).

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 (RKHSs) 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 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 supplementary 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.

To solve this problem, we first substitute $f(x) = Kc$ to approximate $f(x)$ in our hypothesis space H.[7] 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_is, which itself is a possible measure of complexity, but it is weighted to reflect overlap that occurs for points nearer to each other. The resulting 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 supplementary 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 provides the solution to the Tikhonov regularization problem within our flexible space of functions. This estimator is numerically rather benign. Given a fixed value for λ, we compute the kernel matrix and add λ to its diagonal. The resulting matrix is symmetric and positive definite, 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 LM*

The above interpretations informally 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. Here we provide a formal justification for the KRLS approach that offers perhaps less intuition, but has the benefit of being generalizable to other choices of kernels and motivates 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, the mapping $\phi(x_i)$ is infinite-dimensional. Suppose we wish to fit a regularized LM (i.e., a ridge regression) in the expanded features; that is, $f(x_i) = \phi(x_i)^T \theta$, where $\phi(x)$ has dimension D' (which is ∞ in the Gaussian case), and θ is a D' vector of coefficients. Then, we solve

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

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

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.[8]

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 $\langle\theta, \theta\rangle = \left\langle \sum_{i=1}^N c_i\phi(x_i), \sum_{i=1}^N c_i\phi(x_i)\right\rangle = c^TKc$. Thus, both the choice of our function space and our norm can be derived from a ridge regression in a 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 of the KRLS approach and discuss the quantities of interest that can be computed from the KRLS model.

3.1 *Why Gaussian Kernels?*

Although 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 suitable choice. One intuition for this is that we can imagine some data-generating process that produces xs with normally distributed errors. Some xs 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.[9]

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 the mean of y is equivalent to including an (unpenalized) intercept and simplifies the mathematics and exposition. Subtracting the means of

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

[9] 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 $k(x_i, x_j)$.

the xs has no effect, since the kernel is translation-invariant. The rescaling operation is commonly invoked in penalized regressions for norms L_q with $q > 0$—including ridge, bridge, Least Absolute Shrinkage and Selection Operator (LASSO), and elastic-net methods—because, in these approaches, the penalty depends on the magnitudes of the coefficients and thus on the scale of the data. Rescaling by the standard deviation ensures that unit-of-measure decisions have no effect on the estimates. As a second benefit, rescaling enables us to use a simple and fast approach for choosing σ^2 (see below). Note that this rescaling does not interfere with interpretation or generalizability; all estimates are returned to the original scale and location.[10]

3.3 *Choosing the Regularization Parameter* λ

As formulated, there is no single "correct" choice of λ, a property shared with other penalized regression approaches such as ridge, bridge, LASSO, etc. Nevertheless, cross-validation provides a now standard approach (see, e.g., Hastie, Tibshirani, and Friedman 2009) for choosing reasonable values that perform well in practice. We follow the previous work on RLS-related approaches and choose λ by minimizing the sum of the squared leave-one-out errors (LOOEs) by default (e.g., Schölkopf and Smola 2002; Rifkin, Yeo, and Poggio 2003; Rifkin and Lippert 2007). For leave-one-out validation, the model is trained on $N-1$ observations and tested on the left-out observation. For a given test value of λ, this can be done N times, producing a prediction for each observation that does not depend on that observation itself. The N errors from these predictions can then be summed and squared to measure the goodness of out-of-sample fit for that choice of λ. Fortunately, with KRLS, the vector of N LOOEs can be efficiently estimated in $O(N^1)$ time for any valid choice of λ using the formula $\text{LOOE} = \frac{c}{\text{diag}(G^{-1})}$, where $G = K + \lambda I$ (see Rifkin and Lippert 2007).[11]

3.4 *Choosing the Kernel Bandwidth* σ^2

To avoid confusion, we first emphasize that the role of σ^2 in KRLS differs from its role in methods such as traditional kernel regression and kernel density estimation. In those approaches, the kernel bandwidth is typically the only smoothing parameter; no additional fitting procedure is conducted to minimize an objective function, and no separate complexity penalty is available. In KRLS, by contrast, the kernel is used to form K, beyond which fitting is conducted through the choice of coefficients c, under a penalty for complexity controlled by λ. Here, σ^2 enters principally as a measurement decision incorporated into the kernel definition, determining how distant points need to be in the (standardized) covariate space before they are considered dissimilar. The resulting fit is thus expected to be less dependent on the exact choice of σ^2 than is true of those kernel methods in which the bandwidth is the only parameter. Moreover, since there is a trade-off between σ^2 and λ (increasing either can increase smoothness), a range of σ^2 values is typically acceptable and leads to similar fits after optimizing over λ.

Accordingly, in KRLS, our goal is to chose σ^2 to ensure that the columns of K carry useful information extracted from X, resulting in some units being considered similar, some being dissimilar, and some in between. We propose that $\sigma^2 = \dim(X) = D$ is a suitable default choice that adds no computational cost. The theoretical motivation for this proposition is that, in the 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 supplementary appendix). Choosing σ^2 to be

[10] New test points for which estimates are required can be applied, using the means and standard deviations from the original training. Our companion software handles this automatically.

[11] A variant on this approach, generalized cross-validation (GCV), is equal to a weighted version of LOOE (Golub, Heath, and Wahba 1979), computed as $\frac{c}{\frac{1}{N}tr(G^{-1})}$. GCV can provide computational savings in some contexts (since the trace of G^{-1} can be computed without computing G^{-1} itself) but less so here, as we must compute G^{-1} anyway to solve for c. In practice, LOOE and GCV provide nearly identical measures of out-of-sample fit, and commonly, very similar results. Our companion software also allows users to set their own value of λ, which can be used to implement other approaches if needed.

proportional to D therefore ensures a reasonable scaling of the average distance. Empirically, we have found that setting $\sigma^2 = 1D$ in particular has reliably resulted in good empirical performance (see simulations below) and typically provides a suitable distribution of values in K such that entries range from close to 1 (highly similar) to close to 0 (highly dissimilar), with a distribution falling in between.[12]

4 Inference and Interpretation with KRLS

In this section, we provide the properties of the KRLS estimator. In particular, we establish its unbiasedness, consistency, and asymptotic normality and derive a closed-form estimator for its variance.[13] We also develop new interpretational tools, including estimators for the pointwise partial derivatives and their variances, and discuss how the KRLS estimator protects against extrapolation when modeling extreme counterfactuals.

4.1 *Unbiasedness, Variance, Consistency, and Asymptotic Normality*

4.1.1 Unbiasedness

We first show that KRLS unbiasedly estimates the best approximation of the true conditional expectation function that falls in the available space of functions given 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, as we saw above, the functional form assumption in KRLS is much more flexible compared to linear regression or GLMs more generally, and this guards against misspecification bias.

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

This assumption is mathematically equivalent to the usual zero conditional mean assumption used to establish unbiasedness for linear regression or GLMs more generally. However, note that substantively, this assumption is typically weaker in KRLS than in GLMs, which is the source of KRLS' improved robustness to misspecification bias. In a standard OLS setup, with $y = X\beta + \epsilon_{\text{linear}}$, unbiasedness requires that $E[\epsilon_{\text{linear}}|X] = 0$. Importantly, this ϵ_{linear} includes both omitted variables *and* unmodeled effects of X on y that are not linear functions of X (e.g., an omitted squared term or interaction). Thus, in addition to any omitted-variable bias due to

[12] Note that our choice for σ is consistent with advice from other work. For example, Schölkopf and Smola (2002) suggest that an "educated guess" for σ^2 can be made by ensuring that $\frac{(x_i - x_j)^2}{\sigma^2}$ "roughly lies in the same range, even if the scaling and dimension of the data are different," and they also choose $\sigma^2 = \dim(X)$ for the Gaussian kernel in several examples (though without the justification given here). Our companion software also allows users to set their own value for σ^2, and this feature can be used to implement more complicated approaches if needed. In principle, one could also use a joint grid search over values of σ^2 and λ, for example using k-fold cross-validation, where k is typically between five and ten. However, this approach adds a significant computational burden (since a new K needs to be formed for each choice of σ^2), and the benefits can be small since σ^2 and λ trade off with each other, and so it is typically computationally more efficient to fix σ^2 at a reasonable value and optimize over λ.

[13] Although 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, Caponnetto, and Rosasco 2005).

unobserved confounders, misspecification bias also occurs whenever the unmodeled effects of X in ϵ_{linear} are correlated with the Xs that are included in the model. In KRLS, we instead have $y = Kc + \epsilon_{\text{KRLS}}$. In this case, ϵ_{KRLS} is devoid of virtually any smooth function of X because these functions are captured in the flexible model through Kc. In other words, KRLS moves many otherwise unmodeled effects of X from the error term into the model. This greatly reduces the chances of misspecification bias, leaving the errors restricted to principally the unobserved confounders, which will always be an issue in nonexperimental data.

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 this unbiasedness result, 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 supplementary appendix.*

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

We emphasize that this definition of unbiasedness says only that the estimator is unbiased for the best approximation of the conditional expectation function given penalization.[14] 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. Although this may seem like a somewhat dissatisfying notion of unbiasedness, it is precisely the sense in which many other approaches are unbiased, including OLS. 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 an LM will not correctly fit a nonlinear 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, $c^\star$, conditional on a given λ. 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. We note that, as in OLS, the values for the point estimates of interest (e.g., $\hat{y}$, $\frac{\partial \hat{y}}{\partial x^{(d)}}$, discussed below) do not depend on this homoscedasticity assumption. Rather, an assumption over the error structure is needed for computing variances.

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 $\text{Var}[\hat{c}^\star|X,\lambda] = \sigma_\epsilon^2(K+\lambda I)^{-2}$. *The proof is given in the supplementary appendix.*

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

[14] Readers will recognize that classical ridge regression, usually in the span of X rather than $\phi(X)$, 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 postpenalization function because regularization is desirable to identify the most reliable function for making new predictions, the procedure is unbiased for estimating that postpenalization function.

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 $\hat{y}_{\text{test}} = K_{\text{test}}\hat{c}^{\star}$, where K_{test} is the $N_{\text{test}} \times N_{\text{train}}$ dimensional kernel matrix that contains the similarity measures of each test observation to each training observation.[15]

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

Our companion software implements these variance estimators. We estimate σ_{ϵ}^{2} by $\hat{\sigma}_{\epsilon}^{2} = \frac{1}{N}\sum_{i}^{N}\hat{\epsilon}^{2} = \frac{1}{N}(y - K\hat{c}^{\star})^{T}(y - K\hat{c}^{\star})$. Note that all variance estimates above are conditional on the user's choice of λ. This is important, since the variance does indeed depend on λ: higher choices of λ always imply the choice of a more stable (but less well-fitting) solution, producing lower variance. Recall that λ is not a random variable with a distribution but, rather, a choice regarding the trade-off of fit and complexity made by the investigator. LOOE provides a reasonable criterion for choosing this parameter, and so variance estimates are given for $\lambda = \lambda_{\text{LOOE}}$.[16]

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, Caponnetto, and Rosasco 2005). Here, we are instead interested in consistency in the classical sense; that is, determining whether $\operatorname{plim}_{N\to\infty}\hat{y}_{i,N} = y_{i}^{\star}$ 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 (1) $\lambda > 0$ *and (2) 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* $\operatorname{plim}_{N\to\infty}\text{Var}[\hat{y}|X,\lambda] = 0$*, so the estimator is therefore consistent with* $\operatorname{plim}_{N\to\infty}\hat{y}_{i,N} = y_{i}^{\star}$ *for all i. The proof is provided in the supplementary 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 nonzero 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 this quantity grows exactly as N does. Without regularization ($\lambda = 0$), new observations translate into added complexity rather than added certainty; 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 it gains from the efficiency boost when simple models are sufficient.

[15] 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}).

[16] Though we suppress the notation, variance estimates are technically conditional on the choice of σ^{2} as well. Recall that, in our setup, σ^{2} is not a random variable; it is set to the dimension of the input data as a mechanical means of rescaling Euclidian distances appropriately.

As we show below, the regularization can rescue so much efficiency that the resulting KRLS model is not much less efficient than an OLS regression even for linear data.

4.1.4 Finite sample and asymptotic distribution of $\hat{y}$

Here, we establish the 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, $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma_\epsilon^2)$.

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 supplementary appendix.*

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

Assumption 6 (Regularity Conditions II).
Let (1) the errors be independently drawn from a distribution with a finite mean and variance and (2) 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; that is, 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 supplementary appendix. The resulting asymptotic distribution used for inference on any given* $\hat{y}_i$ *is*

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

Theorem 4 is corroborated by simulations, which show that 95% confidence intervals based on standard errors computed by this method (1) closely match confidence intervals constructed from a nonparametric bootstrap, and (2) have accurate empirical coverage rates under repeated sampling where new noise vectors are drawn for each iteration.

Taken together, these new results establish the desirable theoretical properties of the KRLS estimator for the conditional expectation: it is unbiased for the best-fitting approximation to the true Conditional Expectation Function (CEF) in a large space of (penalized) functions (Theorems 1 and 2); it is consistent (Theorem 3); and it is 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 data set 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 data set 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 estimated by

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

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 estimate the sample-average partial derivative of y with respect to $x^{(d)}$ at each observation as

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

We also derive the variance of this quantity, and our companion 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 $\hat{\beta}$ 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 it is highly susceptible to misspecification bias, in which the unmodeled 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 and 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 distribution of $\frac{\partial y}{\partial x^{(d)}}$ 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 $\hat{\beta}$ coefficient) would return a result near zero, this would obscure the fact that the variable in question is having a strong effect but in opposite directions depending on the levels 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 $x^{(b)} = 1$.[17] 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 data set in which all Xs retain their original value and all $x^{(b)} = 1$ and then subtracting from this the mean $\hat{y}$ in a data set where all the values of $x^{(b)} = 0$. In the supplementary 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 preprocessing. 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 LMs, for example, counterfactuals are modeled as though the linear trajectory of the CEF 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 simulation examples of KRLS that illustrate certain aspects of its behavior. Further examples are presented in the supplementary 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 due 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 $x \sim \text{Unif}(0, 1)$ and $\epsilon \sim N(0, 0.3)$. 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 Fig. 2A, this single bad leverage point strongly biases the OLS estimates of the average marginal effect downward (open circles), whereas the estimates of the average marginal effect from KRLS are robust even at small sample sizes (closed circles).

[17] 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 zero to one on this variable.

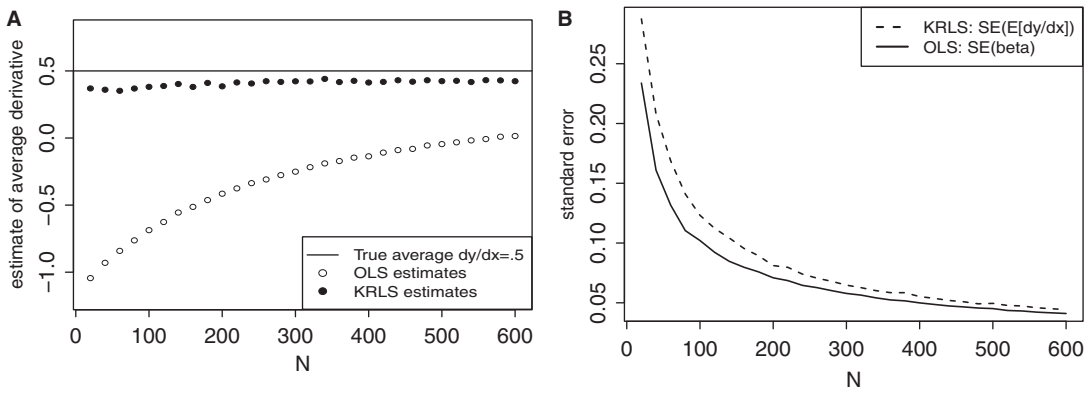

Fig. 2 KRLS compares well to OLS with linear data-generating processes. (A) Simulation to recover the average derivative of $y = 0.5x$; that is, $\frac{\partial y}{\partial x} = 0.5$ (solid line). For each sample size, we run one hundred simulations with observed outcomes $y = 0.5x + \varepsilon$, where $x \sim \text{Unif}(0, 1)$ and $\varepsilon \sim N(0, 0.3)$. 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, whereas OLS is not. (B) Comparison of the 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 $x \sim N(0, 1)$ and $\epsilon \sim N(0, 1)$ with one hundred simulations for each sample size. KRLS is nearly as efficient as OLS at all but very small sample sizes, with standard errors, on average, approximately 14% larger than those of OLS.

5.2 *Efficiency Comparison*

We expect that the added flexibility of KRLS will reduce the bias due to misspecification error but at the cost of increased variance due to the usual bias-variance trade-off. However, 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 $\epsilon \sim N(0, 1)$, such that OLS is guaranteed to be the most efficient unbiased linear estimator according to the Gauss-Markov theorem. Figure 2B 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 only 14% 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 results.

5.3 *Over-Fitting*

A possible concern with flexible estimators is that they may be prone to over-fitting, 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 $y = 0.2\sin(12\pi x) + \sin(2\pi x)$ and run simulations with $x \sim \text{Unif}(0, 1)$ and $\epsilon \sim N(0, 0.2)$ with two sample sizes, $N = 40$ and $N = 400$. The results are displayed in Fig. 3A. 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 LOOE, as described above.

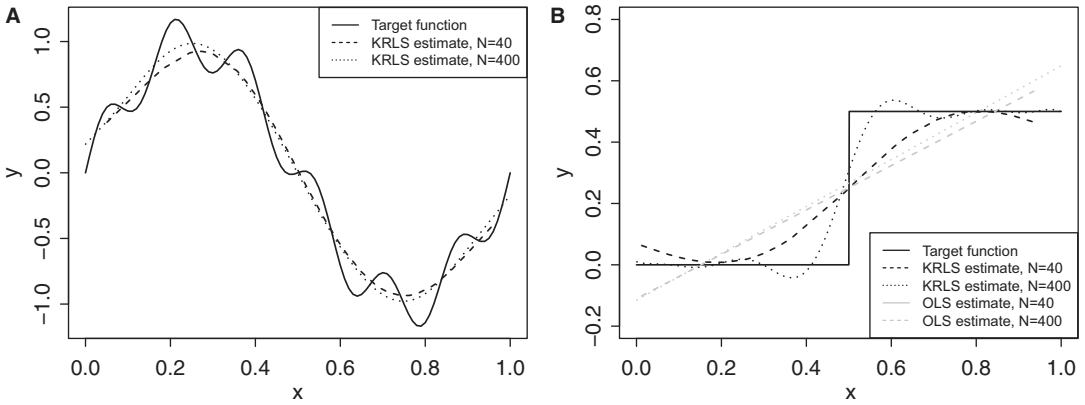

Fig. 3 KRLS with high-frequency and discontinuous functions. (A) Simulation to recover a high-frequency target function given by $y = 0.2 * \sin(12\pi x) + \sin(2\pi x)$ (solid line). For each sample size, we run one hundred simulations where we draw $x \sim \text{Unif}(0, 1)$ and simulate observed outcomes as $y = 0.2 * \sin(12\pi x) + \sin(2\pi x) + \varepsilon$, where $\varepsilon \sim N(0, 0.2)$. The dashed line shows mean estimates across simulations for $N = 40$, and the dotted line for $N = 400$. The results show that KRLS finds a low-frequency approximation even at the larger sample sizes. (B) Simulation to recover the discontinuous target function given by $y = 0.5 * \mathbf{1}(x > 0.5)$ (solid line). For each sample size, we run one hundred simulations where we draw $x \sim \text{Unif}(0, 1)$ and simulate observed outcomes as $y = 0.5 * \mathbf{1}(x > 0.5) + \varepsilon$, where $\varepsilon \sim N(0, 0.2)$. Dashed lines show mean estimates across simulations for $N = 40$, and dotted lines for $N = 400$. The 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.

5.4 *Non-Smooth Functions*

One potential downside of regularization is that KRLS is not well-suited to estimate discontinuous target functions. In Fig. 3B, 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 $x' = \mathbf{1}[x > 0.5]$ 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 $y = 0.5 + x_1 + x_2 - 2(x_1 \cdot x_2) + \varepsilon$ with $x_j \sim \text{Bernoulli}(0.5)$ for $j = 1, 2$ and $\varepsilon \sim N(0, 0.5)$. 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 a thousand test points drawn from the same distribution as the covariates. Figure 4A 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 can 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

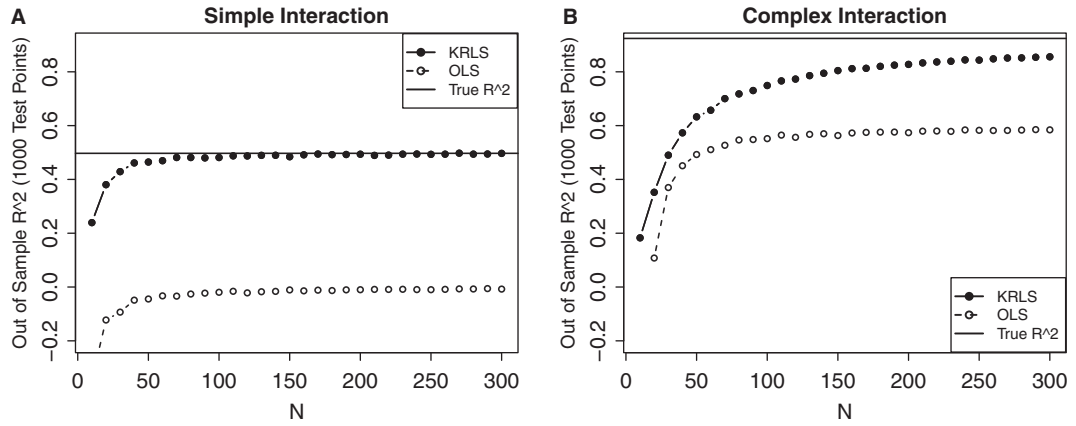

Fig. 4 KRLS learns interactions from the data. Simulations to recover target functions that include multiplicative interaction terms. (A) The target function is $y = 0.5 + x_1 + x_2 - 2(x_1 \cdot x_2) + \varepsilon$ with $x_j \sim$ Bernoulli(0.5) for $j = 1, 2$ and $\varepsilon \sim N(0, 0.5)$. (B) The 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 $p = 0.25$ for x_1 and x_2, $p = 0.75$ for x_3 and x_4, and $p = 0.5$ for all others. For each sample size, we run one hundred simulations where we draw the x and simulate outcomes using $y = y_{\text{true}} + \varepsilon$, where $\varepsilon \sim N(0,0.5)$ for the training data. We use one thousand 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 that one would obtain knowing the functional form as the sample size increases.

that 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 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, Clark, and Golder 2006). KRLS allows researchers to avoid this problem since it learns the interactions from the data.

Figure 4B 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 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). 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 nonlinearities 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 $x_1 \sim \text{Unif}(0, 2)$ 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

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 (0.40)	9.22 (0.52)	5.22 (0.82)	9.38 (0.85)	14.03 (0.79)
x_2	−1.28 (0.21)	0.02 (0.13)	−0.08 (0.19)	0.00 (0.16)	0.10 (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 parentheses. For KRLS, the table shows the average and the quartiles of the distribution of the pointwise marginal effects. The true target function is $y = 5x_1^2$ and simulated using $y\prime = 5x_1^2 + \varepsilon$ with $\varepsilon \sim (0, 2)$, $x_1 \sim \text{Unif}(0, 2)$, 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 vary from about five at the first quartile to about fourteen at the third quartile.

nonlinearity. This nonlinearity is so mild that, in reasonably noisy samples, even a careful researcher who follows the textbook recommendations and first inspects a scatterplot between the outcome and x_1 might 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 who 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 $R^2 = 0.89$. 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 artifact 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 (so 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 of heterogeneity by examining the marginal distribution of the pointwise derivatives. The next three columns display the first, second, and third 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 first quartile to more than 14 at the third quartile. This accurately captures the nonlinearity in the true effect of x_1.

5.7 *Common Interactions and Nonadditivity*

Here, we show how KRLS is well-suited to fit target functions that are nonadditive 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

Table 2 KRLS captures complex interactions and nonadditivity

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 two hundred test points) for simulations using the three target functions displayed in Figs. A4, A5, and A6 in the supplementary appendix with the OLS, GAM, and KRLS estimators. KRLS attains the best in-sample and out-of-sample fit for all three functions.

"valley," two hills and two valleys, or three hills and three valleys (see supplementary appendix, Figs. A4, A5, and A5, 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 two hundred observations, $x_1, x_2 \sim \text{Unif}(0, 1)$, and noise given by $\varepsilon \sim N(0, 0.25)$. We then fit these data using KRLS, OLS, and GAMs. The results are averaged over one hundred simulations. In the supplementary 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 two hundred 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 that one would obtain knowing the functional form. These simulations increase our confidence that KRLS can capture complex nonlinearity, nonadditivity, and interactions that we may expect in social science data. Although 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 nonlinearities 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, because of (1) the appropriateness of its assumptions to social science data, (2) its ease of use, and (3) 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 NNs, regression trees, *k*-nearest neighbors, SVMs, and GAMs, to name a few. In the supplementary appendix, we describe in greater detail how KRLS compares to important classes of models on interpretability and inference, with special attention to 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 misspecification bias than highly constrained models, while minimizing arbitrary choices about basis expansions and the selection of smoothing parameters.

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 two hundred iterations. The data-generating process is based on Wood (2003): $x_1, x_2 \sim \text{Unif}(0,1)$, $\epsilon \sim N(0, 0.25)$, and $y = e^{10(-(x_1-0.25)^2-(x_2-0.25)^2)} + 0.5 * e^{14(-(x_1-0.7)^2-(x_2-0.7)^2)} + \epsilon$. The models are KRLS with default choices; a "naive" GAM (GAM1) that smooths x_1 and x_2 separately; a "smart" GAM (GAM2) that smooths x_1 and x_2 together; a generous 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 an NN with five hidden units. The models are trained on samples of fifty, one hundred, or two hundred observations and then tested on one hundred out-of-sample observations. KRLS outperforms all other methods in small samples. In larger samples, KRLS and the GAM2 (with "full-smoothing") perform similarly. The LM, 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.

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.[18] 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-generating process is given by $x_1, x_2 \sim \text{Unif}(0,1)$, $\epsilon \sim N(0, 0.25)$, and $y = e^{10(-(x_1-0.25)^2-(x_2-0.25)^2)} + 0.5 * e^{14(-(x_1-0.7)^2-(x_2-0.7)^2)} + \epsilon$. 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 smooths x_1 and x_2 separately but then assumes that they add, (3) a "smart" GAM (GAM2) that smooths 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 flexibly specified 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) an NN with five hidden units and all other parameters at their defaults using the NeuralNet package in R. We train this model on samples of fifty, one hundred, or two hundred observations and then test it on one hundred out-of-sample observations. The results for the root mean square error (RMSE) of each model averaged over two hundred 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.[19]

6 Empirical Applications

In this section, we show an application of KRLS to a real data example. In the supplementary appendix, we also provide a second empirical example that shows how KRLS analysis corrects for misspecification bias in a linear interaction model used by Brambor, Clark, and Golder (2006) to test the "short-coattails" hypothesis. This second example highlights the common problem that multiplicative interaction terms in LMs only allow marginal effects to vary linearly, whereas KRLS allows marginal effects to vary in virtually any smooth way, and this added flexibility can be critical to substantive inferences.

[18] It has been shown that the RLS models on which KRLS is based are effective even when used for classification rather than regression, with performance indistinguishable from state-of-the-art Support Vector Machines (Rifkin, Yeo, and Poggio 2003).

[19] KRLS and GAMs in which all variables are smoothed together are 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 supplementary appendix.

6.1 *Predicting Genocide*

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.[20] Harff proposes a "structural model of genocide" where a dummy for genocide onset (*onset*) is regressed on two continuous variables, *prior upheaval* (summed years of prior instability events in the past fifteen years) and *trade openness* (imports and exports as a fraction of gross domestic product [GDP] in logs), and four dummy variables that capture whether the state is an *autocracy*, had a *prior genocide*, and whether the ruling elite has an *ideological character* and/or an *ethnic character*.[21] The first column in Table 4 replicates the original specification, using a linear probability model (LPM) in place of the original logit. We use the LPM here because this allows more direct comparison to the KRLS results. However, the substantive results of the LPM are virtually identical to those of the logit in terms of magnitude and statistical significance. The next four columns on the left present the replication results from the KRLS estimator. We report first differences for all the binary predictor 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 receiver operator curve (ROC) for predicting genocide, with statistically significantly more area under the curve ($p < 0.03$). It is reassuring that KRLS performs better (at least in-sample) than the original logit model even though, as Harff reports, her final specification was selected after an extensive search through a large number of models. Moreover, this added predictive power does not require 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 LPM model (and the original logit) shows a significant effect of *prior upheaval*, with an increase of one standard deviation corresponding to a ten-percentage-point increase in the probability of genocide onset, which corresponds to a 37% increase over the baseline probability. This sizable "effect" 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, although 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 in Table 4 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 marginal effect estimates? It turns out that the significant effect in the LPM model is an artifact of misspecification bias. The variable *prior upheaval* is strongly right-skewed and, when logged to make it more appropriate for linear or logistic regression, the "effect" disappears entirely. This change in results emphasizes the risk of mistaken inference due to misspecification under GLMs and its potential impact on interpretation. Note that this difference in results is by no means trivial substantively. In fact, Harff (2003) argues that *prior upheaval* is "the necessary precondition for genocide and politicide" and "a concept that captures the essence of the structural crises and societal pressures that are preconditions for authorities' efforts to eliminate entire groups." Harff (2003) goes on to explain two mechanisms by which this variable matters and draws policy conclusions from it. However, as the KRLS results show, this "important finding"

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

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

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	0.009*	0.002	−0.001	0.002	0.004
	(0.004)	(0.003)			
Prior genocide	0.263*	0.190*	0.137	0.232	0.266
	(0.119)	(0.075)			
Ideological char. of elite	0.152	0.129	0.086	0.136	0.186
	(0.084)	(0.076)			
Autocracy	0.160*	0.122	0.092	0.114	0.136
	(0.077)	(0.068)			
Ethnic char. of elite	0.120	0.052	0.012	0.046	0.078
	(0.083)	(0.077)			
Trade openness (log)	−0.172*	−0.093*	−0.142	−0.073	−0.048
	(0.057)	(0.035)			
Intercept	0.659				
	(0.217)				

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 parentheses. 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. $N = 126$. *$p < 0.05$.

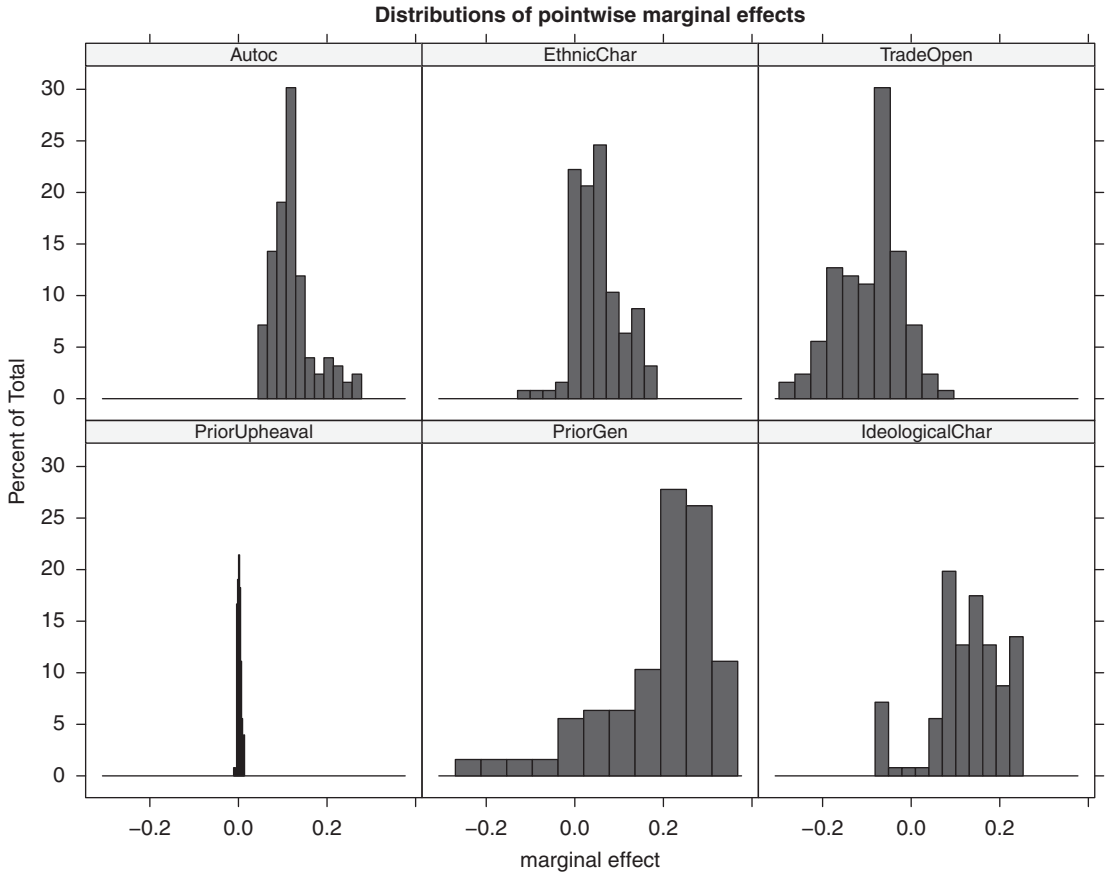

Fig. 5 Effect heterogeneity in Harff data. Histograms of pointwise marginal effects based on KRLS fit to the Harff data (Model 2 in Table 4).

readily disappears when the model accounts for the skew. This showcases the general problem that misspecification bias is often difficult to avoid in typical political science data, even for experienced researchers who 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 coefficient estimates from a regression model and also make richer interpretation possible.

Third, although using KRLS as a robustness test of more rigid models can thus be valuable, working in a much richer model space also permits exploration of effect heterogeneity, including interactions. In Fig. 5 we see that for several variables, such as *autocracy* and *ideological character*, the marginal effect lies to the same side of zero at almost every point, indicating that these variables have marginal effects in the same direction regardless of their level or the levels of other variables. We also see that some variables show little variation in marginal effects, such as *prior upheaval*, whereas others show more substantial variation, such as *prior genocide*.

For example, the marginal effects (measured as first-differences) of *ethnic character* and *ideological character* are mostly positive, but both show variation from approximately zero to twenty percentage points. A suggestive summary of how these marginal effects relate to each observed covariate can be provided by regressing the estimates of the pointwise marginal effects $\frac{\partial \text{ onset}}{\partial \text{ ideological character}}$ or $\frac{\partial \text{onset}}{\partial \text{ ethnic character}}$ on the covariates.[22] Both regressions reveal a strong negative relationship of the level of *trade openness* on these marginal effects. To give substantive interpretation to the results, we find that having an *ethnic character* to the ruling elite is associated with a three-percentage-point higher probability of genocide for countries in the highest quartile of *trade openness*, but a nine-percentage-point higher probability in the highest quartile of *trade openness*. *Ideological character* is associated with a nine-percentage-point higher risk of genocide for the countries in the top quartile of *trade openness*, but an eighteen-percentage-point higher risk among those in the first quartile of *trade openness*. These findings, while associational only, are consistent with theoretical expectations, but would be easily missed in models that do not allow sufficient flexibility.

In addition, the marginal effects of *prior genocide* are very widely dispersed. We find that the marginal effects of *prior genocide* and *ideological character* are strongly related: when one is high, the marginal effect of the other is lessened on average. For example, the marginal effect of *ideological character* is eighteen percentage points higher when *prior genocide* is equal to zero. Correspondingly, the marginal effect of *prior genocide* is twenty-one percentage points higher when *ideological character* is equal to zero. This is characteristic of a sub-additive relationship, in which either prior genocide or ideological character signals a higher risk of genocide, but once one of them is known, the marginal effect of the other is negligible.[23] In contrast, the marginal effects of *ethnic character*—and every other variable besides *ideological character*—changes by little as a function of *prior genocide*.

This brief example demonstrates that KRLS is appropriate and effective in dealing with real-world data even in relatively small data sets. KRLS offers much more flexibility than GLMs and guards against misspecification bias that can result in incorrect substantive

[22]This approach is helpful to identify nonlinearities 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 nonlinearity 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, $x^{(b)}$).

[23]In addition to theoretically plausible reasons why these effects are sub-additive, this relationship may be partly due to ex post facto coding of the variables: once a prior genocide has occurred, it becomes easier to classify a government as having an ideological character, since it has demonstrated a willingness to kill civilians, possibly even stating an ideological aim as justification. Thus, in the absence of *prior genocide*, coding a country as having *ideological character* is informative of genocide risk, whereas it adds less after *prior genocide* has been observed.

inferences. It is also straightforward to interpret the KRLS 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 conveniently generating quantities of interest that are as interpretable and appealing as the coefficients from GLMs. 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 investigators to easily model nonlinear and nonadditive effects and reduce misspecification bias and still produce quantities of interest that enable "simple" interpretations (similar to those allowed by GLMs) and, if desired, more nuanced interpretations that examine nonconstant marginal effects.

Although interpretable quantities can be derived from almost any flexible modeling approach with sufficient knowledge, computational power, and time, constructing such estimates for many methods is inconvenient at best and computationally infeasible in some cases. Moreover, conducting inference over derived quantities of interest multiplies the problem. KRLS belongs to a class of models, those producing continuously differentiable solution surfaces with closed-form expressions, that makes such interpretation feasible and fast. All the interpretational and inferential 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 trade-off 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 is a global method (i.e., the estimate at each point uses information from all other 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 computationally 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 asymptotically normal. Simulations have demonstrated the performance of this method, even with 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 that KRLS is broadly useful whenever investigators are unsure of the functional form in regression and classification problems. This may include model-fitting problems such as prediction tasks, propensity score estimation, or any case where a conditional expectation function must be acquired and rigid functional forms risk missing important variation. The method's interpretability also makes it suitable for both exploratory analyses of marginal effects and causal inference problems in which accurate conditioning on a set of covariates is required to achieve a reliable causal estimate. Relatedly, using KRLS as a specification check for more rigid methods can also be very useful.

However, there remains considerable room for further research. Our hope is that the approach provided here and in our companion software will allow more researchers to begin using KRLS or methods like it; only when tested by a larger community of scholars will we be able to determine the method's true usefulness. Specific research tasks remain as well. Due to the memory demands of working with an $N \times N$ matrix, the practical limit on N for most users is currently in the tens of thousands. Work on 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., G. King, and L. Zeng. 2000. Improving quantitative studies of international conflict: A conjecture. *American Political Science Review* 94:21–36.

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

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

Evgeniou, T., M. Pontil, and T. Poggio. 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* 26(4):797–833.

Golub, G. H., M. Heath, and G. Wahba. 1979. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* 21(2):215–23.

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.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The elements of statistical learning: Data mining, inference, and prediction*. 2nd ed. New York, NY: Springer.

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

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

King, G., and L. Zeng. 2006. The dangers of extreme counterfactuals. *Political Analysis* 14(2):131–59.

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

Rifkin, R., G. Yeo, and T. Poggio. 2003. Regularized least-squares classification. *Nato Science Series Sub Series III Computer and Systems Sciences* 190:131–54.

Saunders, C., A. Gammerman, and V. Vovk. 1998. Ridge regression learning algorithm in dual variables. In *Proceedings of the 15th International Conference on Machine Learning*. Volume 19980, 515–21. San Francisco, CA: Morgan Kaufmann.

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

Tychonoff, A. N. 1963. Solution of incorrectly formulated problems and the regularization method. *Doklady Akademii Nauk SSSR* 151:501–4. Translated in *Soviet Mathematics* 4:1035–8.

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