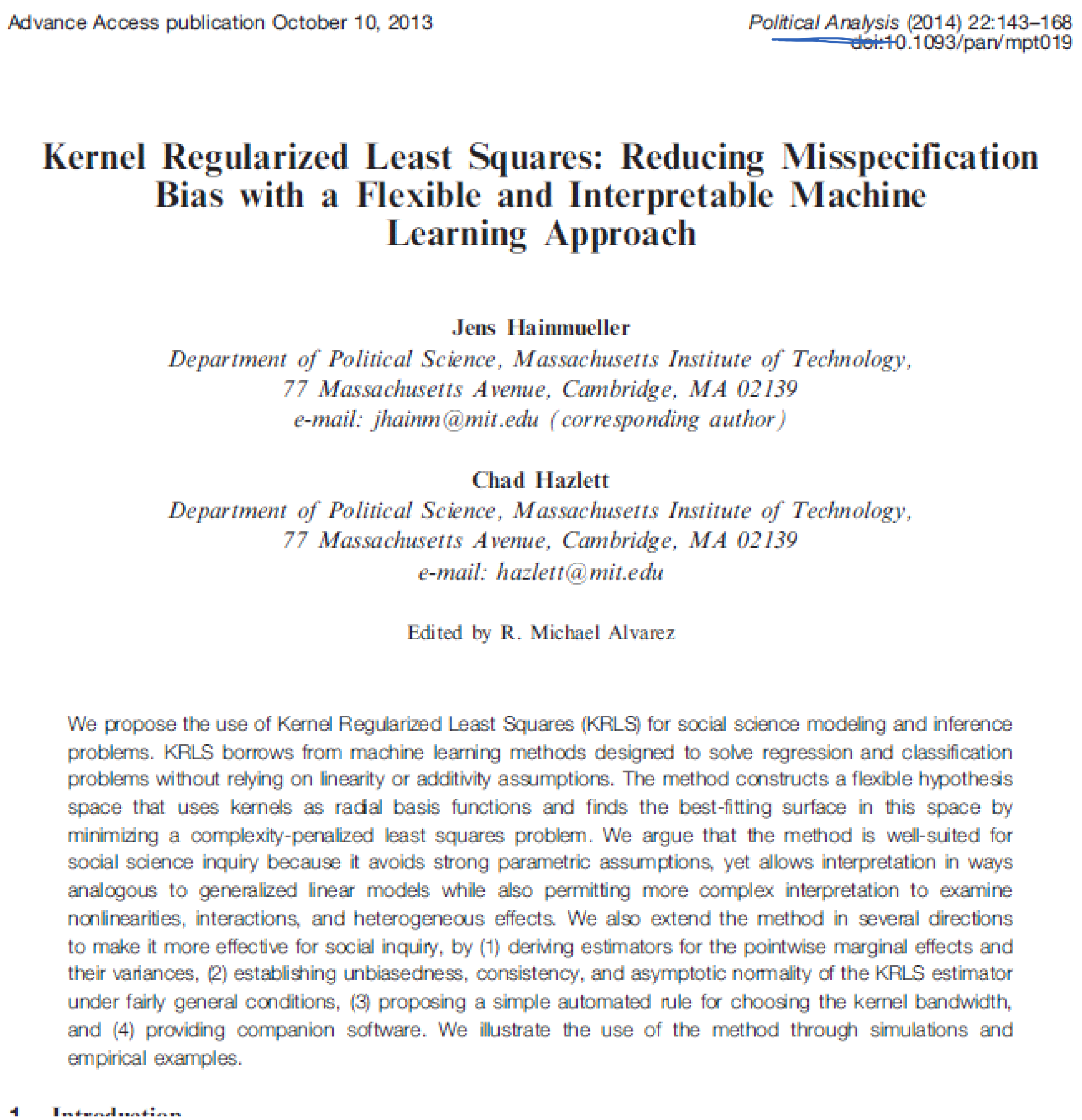
Introduction

Monday, February 20, 2017 5:02 PM



Advance Access publication October 10, 2013

Political Analysis (2014) 22:143–168
~~doi:1~~0.1093/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

Relationship to OLS and Kernel Regression

Monday, February 20, 2017 4:59 PM

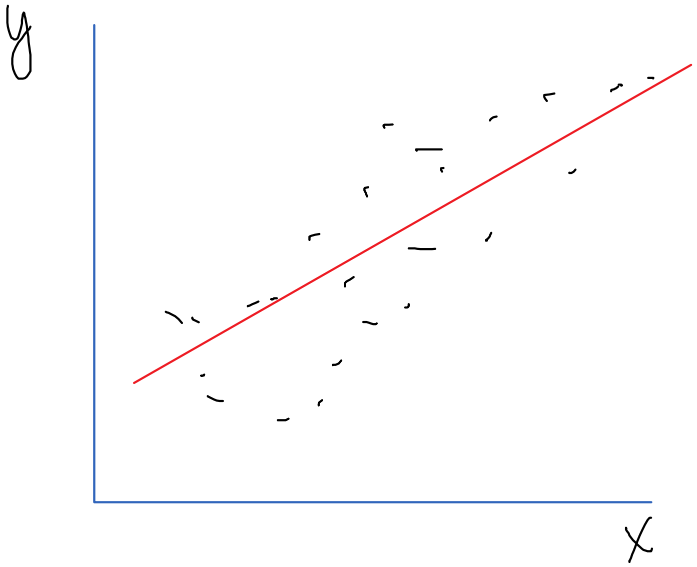
standard linear model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_2 x_k + \varepsilon$$

Matrix notation

$$y = X\underline{\underline{\beta}} + \varepsilon$$

$$\hat{\beta} = (X'X)^{-1}X'y$$



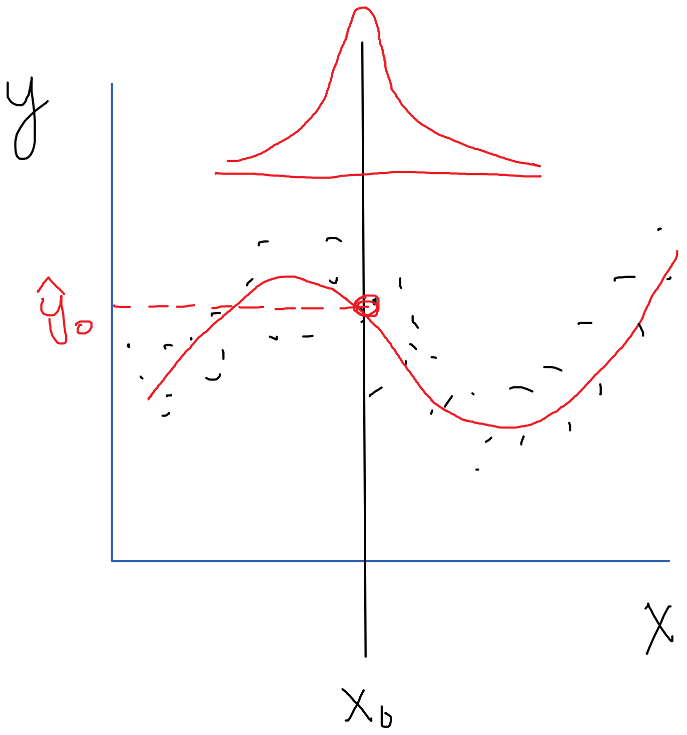
line = good!

kernel regression

degree zero (Nadaraya-Watson)

degree zero (Nadaraya-Watson)

$$\hat{y}_0 = \frac{\underline{\sum_i K(x_i, x_0)\, y_i}}{\sum_i K(x_i, x_0)}$$



KRLS uses two weights:

$\boxed{1}$ constant item coefficient

$\boxed{2}$ kernel distance weight

$\hat{\ } \quad \sum^{N} \alpha \cdot k(x \quad x)$

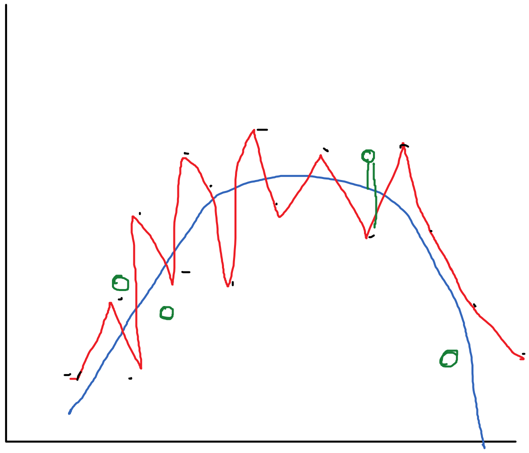
$$\hat{y}_0 = \sum_{i=1}^{N} \underline{\underline{c_i}} * k(x_0, x_i)$$

$$\underset{N\times 1}{\hat{y}} = \underset{N\times N}{K}\,\underset{N\times 1}{\hat{c}} \qquad \text{analogous} \qquad \hat{y} = \underset{N\times k}{X}\,\underset{k\times 1}{\hat{\beta}}$$

$$k(x_0, x_i) = \begin{array}{c|ccccc|} & x_1 & x_2 & x_3 & \cdots & x_n \\ \hline x_1 & - & - & - & & - \\ x_2 & - & - & - & & - \\ x_3 & - & - & - & & - \\ \vdots & & & & & \\ x_n & & & & & \\ \hline \end{array} \qquad N^2$$

Overfitting and Regularization

Monday, February 20, 2017 5:00 PM



$$\arg\min \sum_{i=1}^{N} (\hat{y} - y)^2 \rightarrow \arg\min_{\hat{\beta}} \sum_{i=1}^{N} (X\hat{\beta} - y)^2$$

complexity penalty → $\boxed{\underline{\underline{\lambda}} \|\hat{y}\|^2}$

regularization parameter

$$\|\hat{y}\|^2 = \sum_{i=1}^{N} \hat{y}_i^2$$

→ 0 when KRLS flat.

$$\frac{y - \bar{y}}{\sigma_y} = y_p$$

KRLS problem

$$\arg\min_{\hat{y}} \sum_{i=1}^{N} (\hat{y}_i(x_i) - y_i)^2 + \underline{\lambda \|y\|^2}.$$

how do you choose λ?

cross-validation: good idea, but computationally expensive.

$$\left[\frac{\hat{C}}{\text{diag}(G^{-1})}\right. \quad \text{where} \quad G = \underset{N\times N}{K} + \overset{\text{scalar}}{\underset{1\times 1}{\lambda}}\,\underset{N\times N}{I}$$

choose λ to minimize this quantity.

Computing Marginal Effects

Monday, February 20, 2017 5:02 PM

Partial derivatives for a variable x_d at a single observation j

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

Expected partial derivative for a variable x_d

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

Expected Binary first difference

$$\tfrac{1}{N}\sum[\hat{y}_i | x_i^{(b)} = 1, X = x_i] - \tfrac{1}{N}\sum[\hat{y}_i | x_i^{(b)} = 0, X = x_i]$$