Function Estimation using Reproducing Kernel Hilbert Spaces

Oliver Brown

Estimating Housing Value

- **Goal: Predict Median Home Value using 13 predictor variables and 506 observations**
- Linear Regression? Polynomial Regression?
- What happens when these variables have very complex relationships?
- **What about using Reproducing Kernel Hilbert Spaces?**

Motivation

- How do we estimate an unknown function from data?
- We’ll want a function that balances flexibility and smoothness

What is a Reproducing Kernel Hilbert Space?

- The idea behind an RKHS is to use a space of functions defined by a kernel, so that evaluation and fitting reduce to weighted sums of kernel functions.

Kernels

- Every kernel takes the form of a function K(x, x’)
- A symmetric positive‑semidefinite kernel K is defined by the inner product of two feature‑map vectors:

$$K(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$$

  where each ϕ(x) is a vector (in other words, a list of numbers). The ‘kernel trick’ means we compute this inner product K directly, without ever forming ϕ(x)

- When using our kernel to estimate a function, the solution always takes the form of

$$\sum_{j=1}^{n} \alpha_j \mathcal{K}(x_j, x) = f(x).$$

  where the α coefficients are

Kernel Ridge Regression

- We want a function that fits the data well

$$\min_{f \in \mathbb{H}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

- But we also want to keep the function smooth and cautious of bias
- Kernel Ridge Regression does this by combining two goals:
    - Fit the observed data closely
    - Keep the function smooth and avoidant to overfitting using a penalty

$$\widehat{f} = \arg\min_{f \in \mathbb{H}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda_n \|f\|_{\mathbb{H}}^2 \right\}.$$

- We choose our weights by solving a linear system involving our kernel matrix and a tuning parameter:

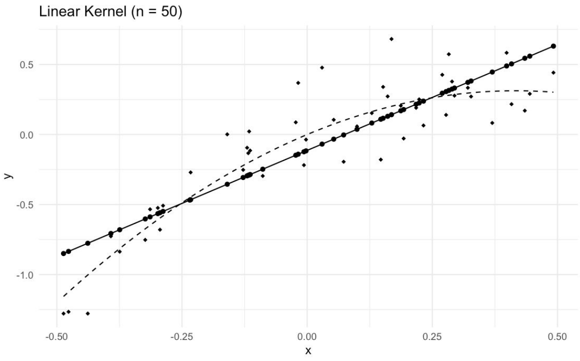
$$\widehat{\alpha} = (\mathbf{K} + \lambda_n \mathbf{I}_n)^{-1} \frac{y}{\sqrt{n}}.$$

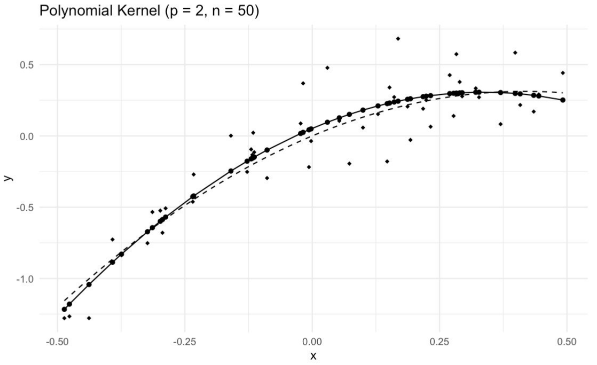
Simulations Introduction

- Goal: Evaluate how different kernels in Kernel Ridge Regression recover known functions under controlled conditions.
- We pick some function and generate noisy observations
  - Our x value is randomly generated from a uniform distribution
  - Our y values is generated using f(x) with a normally distributed random error
- We then apply four different kernels: Linear, Polynomial, Gaussian, and Sobolev to the simulated observations with different sample size
- For each model and sample size, we'll use a 5-fold cross validation to choose a penalty value
- We use the mean squared error on left out data (via 5-fold cross validation) as our metric for how well our estimated function fits the true function

Simulations

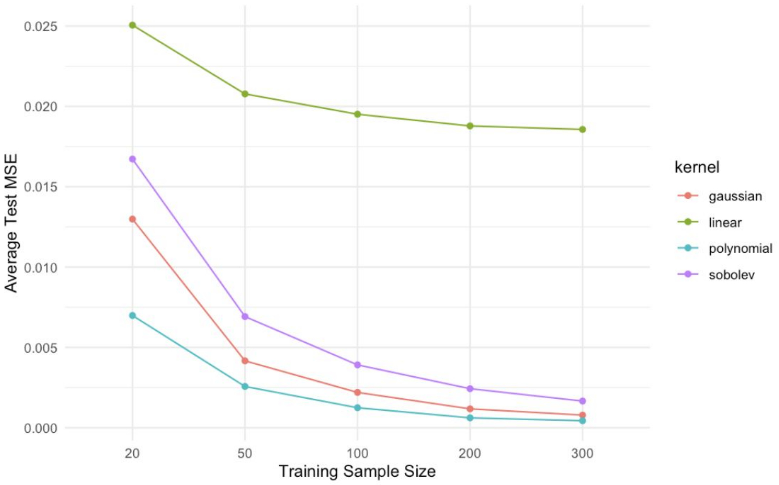
$$f(x) = 3x^2 - \frac{9}{5}x \quad [-0.5,0.5]$$


Linear Kernel (n = 50)


Polynomial Kernel (p = 2, n = 50)

Best to worst:

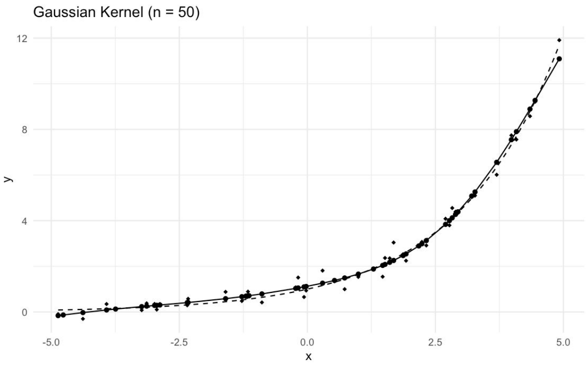
1. Polynomial Model (2nd Order)
2. Gaussian Model (Bandwith ≃ 0.3)
3. Sobolev Model
4. Linear Model

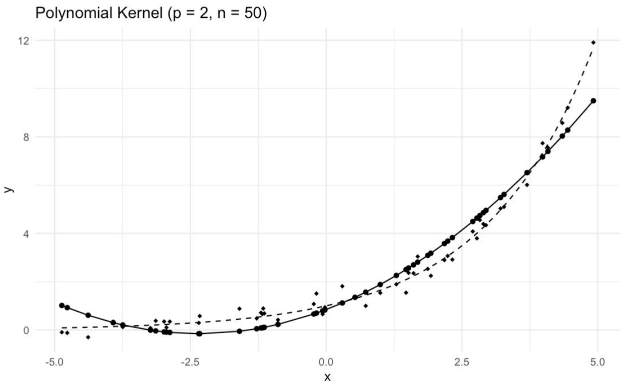


Seed: 1, e □ N(0, 0.2^2), *Bandwidth helps adjust the smoothness of the function

Simulations

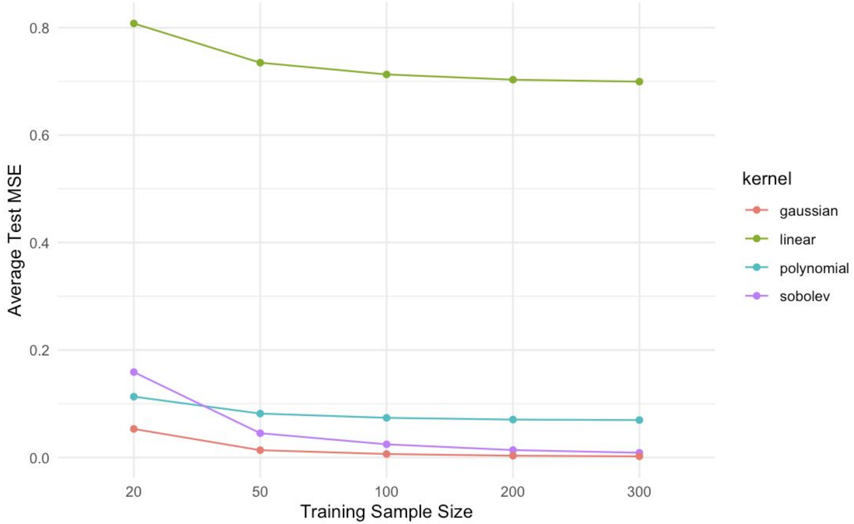
$$f(x) = \exp\left(\frac{x}{2}\right) \quad [-4,4]$$


Gaussian Kernel (n = 50)


Polynomial Kernel (p = 2, n = 50)

Best to worst:

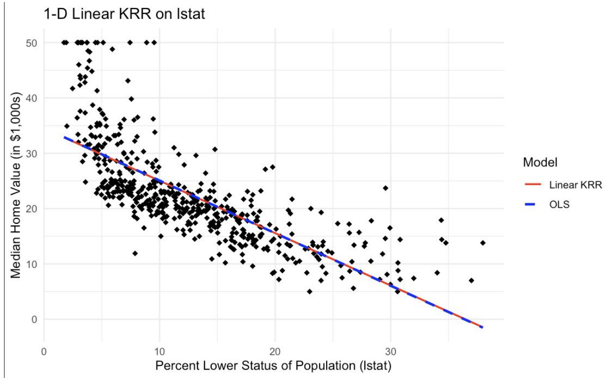
1. Gaussian Model (Bandwith ≃ 3)
2. Sobolev Model
3. Polynomial Model (2nd Order)
4. Linear Model

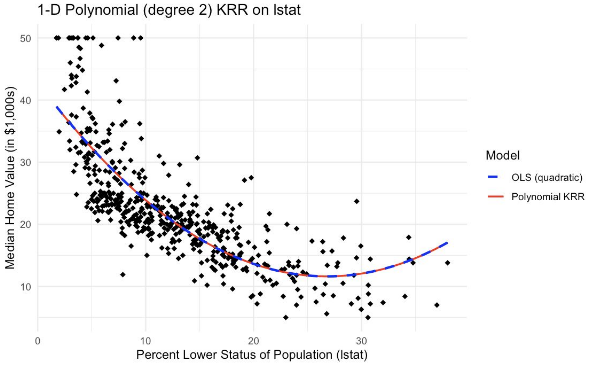


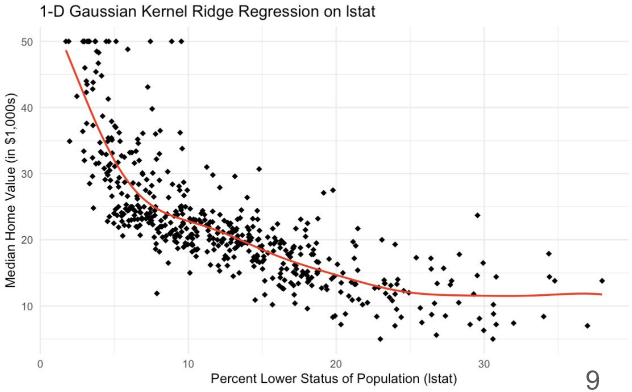
Seed: 1, e □ N(0, 0.3^2)

Application Recap

- Goal: Predict median home value (in $1000s) from 13 features (crime rate, # of rooms, property-tax, accessibility to highways, etc..) from 506 observations
- Method: Kernel Ridge Regression using five different kernel functions
    - Linear, Polynomial, Gaussian, Sobolev, and Cosine







Application (Cont.)

- Using all 13 predictors
- Median distance between x-values was used for bandwidth (gaussian)
- Also considered using linear and polynomial ridge regression to compare to linear and polynomial kernels

Results:

- Linear kernel performed about the same as their linear RR counterpart.
- Polynomial Kernel performed better than their polynomial RR counterpart
- Gaussian Model is the best for predicting median home value for this data
(RMSE = approx. $3770)

Model	**Avg. MSE from 5-fold CV**
Linear KRR	24.52
Linear Ridge Regression	24.14
Polynomial KRR (p = 2) **	16.57
Quadratic Ridge Regression	19.04
Gaussian KRR *	14.25
Sobolev KRR ***	18.38
Cosine KRR	33.80

References

- **Text:** Wainwright MJ. High-Dimensional Statistics: A Non-Asymptotic Viewpoint. Cambridge University Press; 2019.
- **Dataset:** Harrison, D., & Rubinfeld, L. (1978). Hedonic prices and the demand for clean air. Journal of Environmental Economics and Management, 5(1), 81–102.

Packages Used

- ggplot2 (for plots)
- MASS (for bostonhousing dataset)
- glmnet (for linear and polynomial ridge regression models)

Also, thank you Antonio!