

Mean Squared Error in Model Selection

Adam L. Pintar^{1,2}
Christine M. Anderson-Cook²
Huaiqing Wu¹

¹Department of Statistics, Iowa State University

²Statistical Sciences, Los Alamos National Laboratory

June 3, 2009

LA-UR: 09-03286

Outline

Existing Work

Goal

Algorithm

Overview

Covariate Space

bias², Variance, MSE

Compare Models

Standard Results

Real Data Example

Introduction

Covariate Space

Results

Standard Results

Discussion &

Conclusions

References

Outline

- ▶ Some existing work in variable selection
- ▶ Goal of methodology
- ▶ Our algorithm via an example
- ▶ A real data example
- ▶ Discussion and conclusions

Mean Squared
Error in Model
Selection

Pintar,
Anderson-Cook,
Wu

Outline

Existing Work

Goal

Algorithm

Overview

Covariate Space

bias², Variance, MSE

Compare Models

Standard Results

Real Data Example

Introduction

Covariate Space

Results

Standard Results

Discussion &
Conclusions

References

Many Methods of Variable selection

- ▶ Some of the most popular methods
 - ▶ AIC (Akaike (1974))
 - ▶ BIC (Schwarz (1978))
 - ▶ Cross Validation (Shao (1993))
 - ▶ Mallows C_p (Mallows (1973))
 - ▶ Adjusted R^2
 - ▶ Stepwise Selection
 - ▶ Stochastic Search Variable Selection (George and McCulloch (1993))
- ▶ All consider only the observed data

Mean Squared
Error in Model
Selection

Pintar,
Anderson-Cook,
Wu

Outline

Existing Work

Goal

Algorithm

Overview

Covariate Space

bias², Variance, MSE

Compare Models

Standard Results

Real Data Example

Introduction

Covariate Space

Results

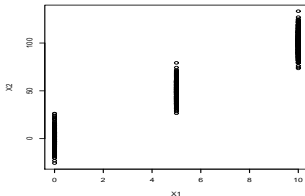
Standard Results

Discussion &
Conclusions

References

Goal

- ▶ Consider probit regression models
- ▶ p - True system reliability
- ▶ \hat{p}^m - Estimated system reliability under model m
- ▶ Provide a variable selection algorithm, focused on prediction, in a user defined region of the covariate space



Mean Squared
Error in Model
Selection

Pintar,
Anderson-Cook,
Wu

Outline

Existing Work

Goal

Algorithm

Overview

Covariate Space

bias², Variance, MSE

Compare Models

Standard Results

Real Data Example

Introduction

Covariate Space

Results

Standard Results

Discussion &
Conclusions

References

Overview of Our Algorithm

- ▶ Select and characterize the user-specified region of interest in the covariate space
- ▶ Randomly sample new locations from the region of interest
- ▶ Estimate prediction bias, prediction variance, and prediction MSE at all sampled locations for all models to be compared
- ▶ Compare models graphically, based on the estimated values in the previous step to select a best model
 - ▶ Focus is on MSE

Mean Squared
Error in Model
Selection

Pintar,
Anderson-Cook,
Wu

Outline

Existing Work

Goal

Algorithm

Overview

Covariate Space

bias², Variance, MSE

Compare Models

Standard Results

Real Data Example

Introduction

Covariate Space

Results

Standard Results

Discussion &
Conclusions

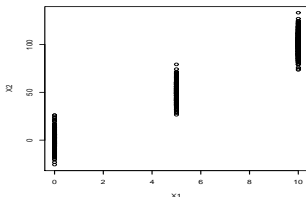
References

Why Focus on MSE?

- ▶ Ideal to simultaneously minimize prediction variance and prediction bias
- ▶ $MSE(\hat{p}^m) = \text{variance}(\hat{p}^m) + \text{bias}(\hat{p}^m)^2$
- ▶ MSE balances variance and bias, which is a compromise to minimizing both
- ▶ One issue
 - ▶ $\text{bias}(\hat{p}^m) = E(\hat{p}^m) - p$
 - ▶ Need a surrogate for p
 - ▶ The estimated reliability from the full model is used

Example 1 Introduction

- ▶ Two covariates X_1 and X_2
 - ▶ $(X_2 | X_1) \sim N(10 * X_1, 100)$



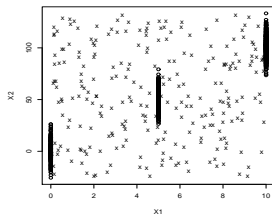
- ▶ The response $Y_i \sim \text{Bernoulli}(p_i)$
 - ▶ $\Phi^{-1}(p_i) = 2.3 - 0.1 * X_1 - 0.02 * X_2$

Selecting a Region of the Covariate Space

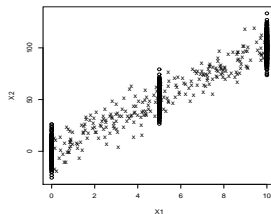
Mean Squared
Error in Model
Selection

Pintar,
Anderson-Cook,
Wu

- ▶ Suppose we are interested in predicting reliability for $X_1 \in [0, 10]$



(a)



(b)

Outline

Existing Work

Goal

Algorithm

Overview

Covariate Space

bias², Variance, MSE

Compare Models

Standard Results

Real Data Example

Introduction

Covariate Space

Results

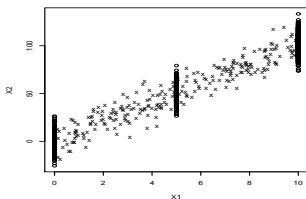
Standard Results

Discussion &
Conclusions

References

Characterizing the Relationship Between Covariates and Sampling

- ▶ Simple linear regression for characterization
- ▶ Sampling
 - ▶ Regress X_2 on X_1 using the observed points.
 - ▶ Sample $X_1 \in [0, 10]$ uniformly.
 - ▶ For every sampled X_1 value, sample $X_2 \sim N(b_0 + b_1 * X_1, \hat{\sigma}^2)$
 - ▶ Here, $b_0 = 0.76$, $b_1 = 9.97$, and $\hat{\sigma}^2 = 97.73$



Calculation Details

- ▶ For each model under consideration, at each sampled point calculate bias², variance, and MSE
- ▶ $\hat{\beta}^f$ - estimated regression coefficients for the full model
- ▶ $\hat{\beta}^m$ - estimated regression coefficients for model m
- ▶ $\hat{p} = \Phi(x' \hat{\beta}^f)$ - estimated true system reliability under the full model at covariate location x
- ▶ $\hat{p}^m = \Phi(x' \hat{\beta}^m)$ - estimated reliability under model m at covariate location x
- ▶ $\hat{Var}(\hat{p}^m) = \left[\left(\frac{\partial p^m}{\partial \beta} \right)' \right]_{\beta^m = \hat{\beta}^m} \hat{\Sigma} \beta^m \left[\left(\frac{\partial p^m}{\partial \beta} \right) \right]_{\beta^m = \hat{\beta}^m}$
- ▶ $\hat{bias}^m = \hat{p}^m - \hat{p}$
- ▶ $\hat{MSE}^m = (\hat{bias}^m)^2 + \hat{Var}(\hat{p}^m)$
- ▶ Note that bias is estimated as zero for the full model

Naming the Models

| Model | X_1 | X_2 | $X_1 * X_2$ |
|-------|-------|-------|-------------|
| 1 | 0 | 0 | 0 |
| 2 | 0 | 0 | 1 |
| 3 | 0 | 1 | 0 |
| 4 | 0 | 1 | 1 |
| 5 | 1 | 0 | 0 |
| 6 | 1 | 0 | 1 |
| 7 | 1 | 1 | 0 |
| 8 | 1 | 1 | 1 |

Mean Squared
Error in Model
Selection

Pintar,
Anderson-Cook,
Wu

Outline

Existing Work

Goal

Algorithm

Overview

Covariate Space

bias², Variance, MSE

Compare Models

Standard Results

Real Data Example

Introduction

Covariate Space

Results

Standard Results

Discussion &
Conclusions

References

Boxplots

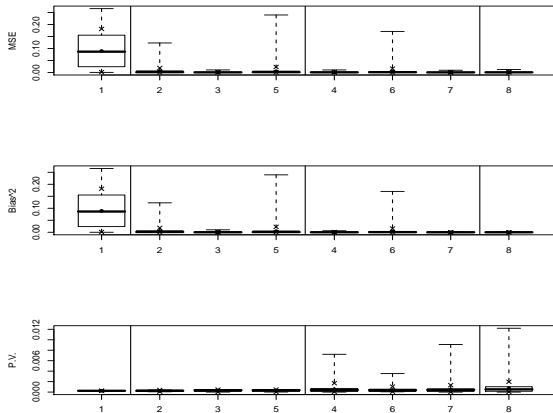


Figure: Boxplots of MSE, bias^2 , and variance

Fraction of Design Space (FDS) Plots

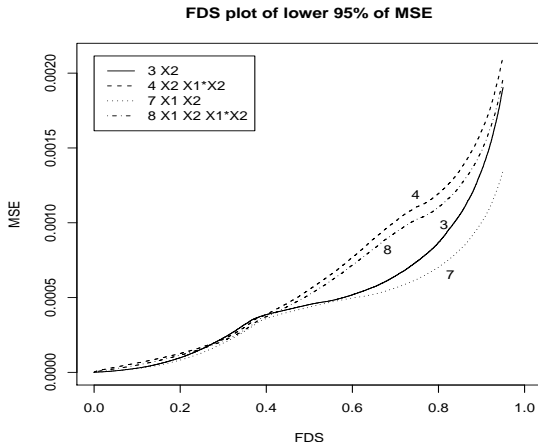


Figure: FDS curves of the four best models with respect to MSE Zahran et al. (2003).

Mean Squared
Error in Model
Selection

Pintar,
Anderson-Cook,
Wu

Outline

Existing Work

Goal

Algorithm

Overview

Covariate Space

bias², Variance, MSE

Compare Models

Standard Results

Real Data Example

Introduction

Covariate Space

Results

Standard Results

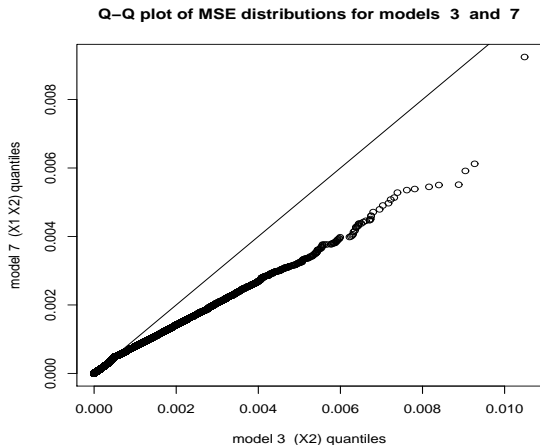
Discussion &
Conclusions

References

Quantile-Quantile Plots

Mean Squared
Error in Model
Selection

Pintar,
Anderson-Cook,
Wu



Outline

Existing Work

Goal

Algorithm

Overview

Covariate Space

bias², Variance, MSE

Compare Models

Standard Results

Real Data Example

Introduction

Covariate Space

Results

Standard Results

Discussion &
Conclusions

References

Results from AIC and Cross Validation

Mean Squared
Error in Model
Selection

Pintar,
Anderson-Cook,
Wu

| Model Name | Model Terms | AIC Value | Cross validation $\hat{\Gamma}$ |
|------------|-----------------------|-----------------|---------------------------------|
| 1 | none | 1369.741 | 0.2470 |
| 2 | $X_1 * X_2$ | 802.9641 | 0.1278 |
| 3 | X_2 | 784.1065 | 0.1261 |
| 4 | $X_2, X_1 * X_2$ | 785.8046 | <i>0.1269</i> |
| 5 | X_1 | 807.8247 | 0.1305 |
| 6 | $X_1, X_1 * X_2$ | 797.5526 | 0.1287 |
| 7 | X_1, X_2 | 784.2477 | 0.1271 |
| 8 | $X_1, X_2, X_1 * X_2$ | 786.2079 | 0.1272 |

- ▶ True model
 - ▶ $\Phi^{-1}(p_i) = 2.3 - 0.1 * X_1 - 0.02 * X_2$
- ▶ Our method identifies correct model
- ▶ Standard methods emphasize a model without X_1
 - ▶ X_1 is only observed at three distinct values

Outline

Existing Work

Goal

Algorithm

Overview

Covariate Space

bias², Variance, MSE

Compare Models

Standard Results

Real Data Example

Introduction

Covariate Space

Results

Standard Results

Discussion &

Conclusions

References

Example 2 Introduction

Mean Squared
Error in Model
Selection

Pintar,
Anderson-Cook,
Wu

- ▶ Background
 - ▶ The responses are pass/fail results collected from a missile system
 - ▶ The available covariates are *age* in years, and *usage* in hours in ready mode
 - ▶ Due to the proprietary nature of the full systems, the actual pass/fail results for individual systems has been adjusted
 - ▶ Use a probit regression model to describe the data
- ▶ Characterizing the relationship between *age* and *usage*
 - ▶ Start with a scatter plot

Outline

Existing Work

Goal

Algorithm

Overview

Covariate Space

bias², Variance, MSE

Compare Models

Standard Results

Real Data Example

Introduction

Covariate Space

Results

Standard Results

Discussion &
Conclusions

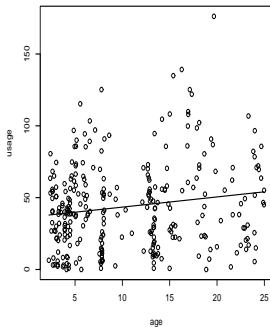
References

The Relationship Between *age* and *usage*

Mean Squared
Error in Model
Selection

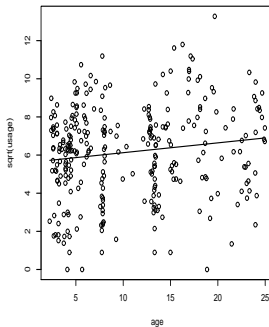
Pintar,
Anderson-Cook,
Wu

Scatter Plot of Age Versus Usage, with Regression Line



(a)

Scatter Plot of Age Versus $\sqrt{\text{Usage}}$, with Regression Line



(b)

Outline

Existing Work

Goal

Algorithm

Overview

Covariate Space

bias^2 , Variance, MSE

Compare Models

Standard Results

Real Data Example

Introduction

Covariate Space

Results

Standard Results

Discussion &
Conclusions

References

Characterizing the Covariate Space, Choosing the Region, and Sampling

- ▶ Linear regression is an appropriate description
- ▶ Specifying a range for *age* describes the region
- ▶ Decisions are made based on the prediction of the future
 - ▶ Extrapolation is required
 - ▶ Scientific and engineering understanding
- ▶ The observed range of *age* is about 2 to 25 years.
- ▶ Suppose interest is in making prediction for *age* in 24 to 30 years
- ▶ Sampling
 - ▶ Sample *age* randomly in 24 to 30 years
 - ▶ Sample \sqrt{usage} according to the linear regression for each sampled value of *age*

Mean Squared
Error in Model
Selection

Pintar,
Anderson-Cook,
Wu

Outline

Existing Work

Goal

Algorithm

Overview

Covariate Space

bias², Variance, MSE

Compare Models

Standard Results

Real Data Example

Introduction

Covariate Space

Results

Standard Results

Discussion &
Conclusions

References

Boxplots

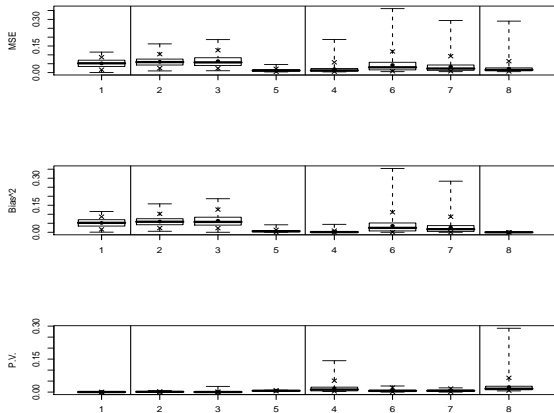


Figure: Boxplots of MSE, bias², and variance.

Mean Squared
Error in Model
Selection

Pintar,
Anderson-Cook,
Wu

Outline

Existing Work

Goal

Algorithm

Overview

Covariate Space

bias², Variance, MSE

Compare Models

Standard Results

Real Data Example

Introduction

Covariate Space

Results

Standard Results

Discussion &
Conclusions

References

Results from AIC and Cross Validation

Mean Squared
Error in Model
Selection

Pintar,
Anderson-Cook,
Wu

| Model Name | Model Terms | AIC Value | Cross validation $\hat{\Gamma}$ |
|------------|-----------------------|-----------------|---------------------------------|
| 1 | none | 181.9922 | 0.0854 |
| 2 | $X_1 * X_2$ | 183.6414 | 0.0864 |
| 3 | X_2 | 170.1779 | 0.0823 |
| 4 | $X_2, X_1 * X_2$ | 156.3253 | 0.0793 |
| 5 | X_1 | 176.2935 | 0.0849 |
| 6 | $X_1, X_1 * X_2$ | 168.5032 | 0.0843 |
| 7 | X_1, X_2 | 159.7092 | 0.0826 |
| 8 | $X_1, X_2, X_1 * X_2$ | 157.9798 | 0.0825 |

- ▶ AIC and cross validation highlight the same best model
 - ▶ These do not consider extrapolation
- ▶ Our method chooses a smaller model
 - ▶ The extra variance caused by including correlated terms overtakes the reduction in bias in the extrapolation case

Outline

Existing Work

Goal

Algorithm

Overview

Covariate Space

bias², Variance, MSE

Compare Models

Standard Results

Real Data Example

Introduction

Covariate Space

Results

Standard Results

Discussion &
Conclusions

References

Discussion & Conclusions

Mean Squared
Error in Model
Selection

Pintar,
Anderson-Cook,
Wu

- ▶ Recap of Algorithm
 - ▶ Select and characterize region of interest
 - ▶ Randomly sample new locations from the region
 - ▶ Calculate MSE, bias², and variance
 - ▶ Compare models graphically
- ▶ How model will be used should influence selection procedure
- ▶ Characterization of the covariate space is key
- ▶ Able to deal with correlation between explanatory variables
- ▶ Extendable to other model forms

Outline

Existing Work

Goal

Algorithm

Overview

Covariate Space

bias², Variance, MSE

Compare Models

Standard Results

Real Data Example

Introduction

Covariate Space

Results

Standard Results

Discussion &
Conclusions

References

Akaike, H. (1974), "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, 19, 716–723.

George, E. I. and McCulloch, R. E. (1993), "Variable Selection Via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 88–889.

Mallows, C. L. (1973), "Some Comments on C_p ," *Technometrics*, 15, 661–675.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Shao, J. (1993), "Linear Model Selection by Cross-Validation," *Journal of the American Statistical Association*, 88, 486–494.

Zahran, A., Anderson-Cook, C. M., and Myers, R. H. (2003), "Fraction of Design Space to Assess Prediction Capability of Response Surface Designs," *Journal of Quality Technology*, 35, 377–386.

Mean Squared
Error in Model
Selection

Pintar,
Anderson-Cook,
Wu

Outline

Existing Work

Goal

Algorithm

Overview

Covariate Space

bias², Variance, MSE

Compare Models

Standard Results

Real Data Example

Introduction

Covariate Space

Results

Standard Results

Discussion &

Conclusions

References