

# 171:290 Model Selection

## Lecture XI: Discrepancy-Based Model Selection Criteria

Joseph E. Cavanaugh

Department of Biostatistics  
Department of Statistics and Actuarial Science  
The University of Iowa

November 3, 2009

## Introduction

- A model selection criterion is often formulated by constructing an approximately unbiased estimator of an *expected discrepancy*, a measure that gauges the separation between the true model and a fitted candidate model.
- Conventional form of model selection criterion:  
goodness-of-fit term + penalty term.
- The natural estimator of the expected discrepancy, the *estimated discrepancy*, corresponds to the goodness-of-fit term in the selection criterion.

## Introduction

- Expected discrepancy: Reflects how well, on average, the fitted approximating model predicts “new” data generated under the true model.
- Estimated discrepancy: Reflects how well the fitted approximating model predicts the data at hand.

## Introduction

- The estimated discrepancy yields an overly optimistic assessment of how effectively the fitted model predicts new data. It therefore serves as a negatively biased estimator of the expected discrepancy.
- Correcting for this bias leads to the penalty term of the selection criterion.

## Introduction

- In this lecture, we outline the general framework for discrepancy-based model selection criteria.
- The collection of criteria based on discrepancy estimation is very large, and includes AIC, AICc, MAIC, TIC,  $C_p$ ,  $MC_p$ , and FPE.
- Much of the subsequent development is from Linhart and Zucchini, 1986, *Model Selection*.

# Introduction

## Outline:

- Model Selection Framework
- Examples of Discrepancies and Resulting Criteria
  - The Kullback discrepancy and AIC
  - The SSE discrepancy and the Gauss criterion,  $C_p$ , and FPE
- Simulation Study

# Framework

- Let the elements of the data vector  $y$  be denoted as  $y_1, y_2, \dots, y_n$ .
- For simplicity, suppose that the  $y_i$  are independent.

# Framework

- Let  $\mathcal{M}$  denote the collection of distribution functions for  $y$ .
- **True** or **generating model**:  $G \in \mathcal{M}$ .
- **Candidate** or **approximating model**:  $F \in \mathcal{M}$ .
- **Discrepancy**: A mapping from  $\mathcal{M} \times \mathcal{M}$  to  $\mathfrak{R}$  which has the property  $d(G, F) \geq d(G, G)$ .
- $d(G, F)$  should increase as the disparity increases between the generating model  $G$  and the approximating model  $F$ .
- $d(G, F)$  need not be a formal metric.

## Framework

- It is generally assumed that  $F$  is parametric, in which case we may write  $F$  as  $F_{\theta_k}$ ,  $\theta_k \in \Theta(k)$ , and use  $d(G, F_{\theta_k})$  to denote the discrepancy.
- Let  $\mathcal{F}(k) = \{F_{\theta_k} \mid \theta_k \in \Theta(k)\}$  denote a  $k$ -dimensional parametric class, i.e., a class of distributions in which the parameter space  $\Theta(k)$  consists of  $k$ -dimensional vectors whose components are functionally independent.

# Framework

- For simplicity, we will consider discrepancies  $d(G, F_{\theta_k})$  of the following form:

$$d(G, F_{\theta_k}) = \sum_{i=1}^n E_G \{ \delta_i(y_i, \theta_k) \}.$$

- $\delta_i(y_i, \theta_k)$  serves as a case-specific disparity measure that assesses how effectively  $y_i$  is predicted by the approximating model  $F_{\theta_k}$ .

# Framework

- **Parameter estimate:**  $\hat{\theta}_k$ .
- **Fitted model:**  $F_{\hat{\theta}_k}$ .
- Often, the estimate  $\hat{\theta}_k$  is obtained as follows:

$$\hat{\theta}_k = \operatorname{argmin}_{\theta_k \in \Theta(k)} \sum_{i=1}^n \delta_i(y_i, \theta_k).$$

- Such an estimate  $\hat{\theta}_k$  is called the **minimum discrepancy estimate (MDE)**.

# Framework

- **Pseudo true parameter:**

$$\bar{\theta}_k = \operatorname{argmin}_{\theta_k \in \Theta(k)} d(G, F_{\theta_k}).$$

- Interpretation: Of all models in  $\mathcal{F}(k)$ , the model  $F_{\bar{\theta}_k}$  provides the best approximation to  $G$  (where model proximity is gauged in terms of  $d(G, F_{\theta_k})$ ).
- **Best approximating model:**  $F_{\bar{\theta}_k}$ .
- Under suitable regularity conditions, the MDE  $\hat{\theta}_k$  is consistent for  $\bar{\theta}_k$ .

## Framework

- **Discrepancy due to approximation:**

$$d(G, F_{\bar{\theta}_k}) = \sum_{i=1}^n \mathbb{E}_G \{ \delta_i(y_i, \bar{\theta}_k) \}.$$

- This discrepancy represents the disparity between the true model  $G$  and the best approximating model  $F_{\bar{\theta}_k}$ .
- If  $G \in \mathcal{F}(k)$ , then  $G = F_{\bar{\theta}_k}$ , and this discrepancy attains its smallest possible value.
- Otherwise, this discrepancy can be decreased by increasing the complexity (i.e., the dimension) of the candidate class  $\mathcal{F}(k)$ .

## Framework

- **Discrepancy due to estimation:**

$$d(F_{\hat{\theta}_k}, F_{\hat{\theta}_k}) = \sum_{i=1}^n E_{F_{\hat{\theta}_k}} \{ \delta_i(y_i, \theta_k) \} |_{\theta_k = \hat{\theta}_k}.$$

- Linhart and Zucchini (p. 11): “It expresses the magnitude of the lack of fit due to sampling variation.”
- For a fixed sample size  $n$ , this discrepancy can be decreased by decreasing the complexity of the candidate class  $\mathcal{F}(k)$ .

## Framework

- **Overall discrepancy:**

$$d(G, F_{\hat{\theta}_k}) = \sum_{i=1}^n E_G \{ \delta_i(y_i, \theta_k) \} |_{\theta_k = \hat{\theta}_k}.$$

- If the discrepancy of interest is a formal metric, then by the triangle inequality

$$d(G, F_{\hat{\theta}_k}) \leq d(G, F_{\bar{\theta}_k}) + d(F_{\bar{\theta}_k}, F_{\hat{\theta}_k}).$$

- A model selection criterion is often formulated by constructing an approximately unbiased estimator of the **expected overall discrepancy**:

$$\Delta(G, k) = E_G \left\{ d(G, F_{\hat{\theta}_k}) \right\}.$$

## Framework

- $\Delta(G, k)$  reflects how well, on average, fitted approximating models having the same structure as  $F_{\hat{\theta}_k}$  predict “new” data generated under the true model  $G$ :

$$\begin{aligned}\Delta(G, k) &= E_G \left\{ d(G, F_{\hat{\theta}_k}) \right\} \\ &= E_G \left\{ \sum_{i=1}^n E_G \{ \delta_i(y_i, \theta_k) \} \mid \theta_k = \hat{\theta}_k \right\}.\end{aligned}$$

- The natural estimator of the expected discrepancy is the **estimated discrepancy**

$$\hat{d}(\hat{\theta}_k) = \sum_{i=1}^n \delta_i(y_i, \hat{\theta}_k).$$

## Framework

- The estimated discrepancy  $\hat{d}(\hat{\theta}_k)$  reflects how well the fitted approximating model  $F_{\hat{\theta}_k}$  predicts the data at hand,  $y$ .
- $\hat{d}(\hat{\theta}_k)$  yields an overly optimistic assessment of how effectively the fitted model predicts new data.
- $\hat{d}(\hat{\theta}_k)$  therefore serves as a negatively biased estimator of the expected discrepancy  $\Delta(G, k)$ .
- Correcting for this bias leads to the penalty term of the selection criterion.

## Framework

- Consider writing  $\Delta(G, k)$  as follows:

$$\begin{aligned}\Delta(G, k) &= E_G \left\{ d(G, F_{\hat{\theta}_k}) \right\} \\ &= E_G \left\{ \hat{d}(\hat{\theta}_k) \right\} + \\ &\quad \left[ E_G \left\{ d(G, F_{\hat{\theta}_k}) - \hat{d}(\hat{\theta}_k) \right\} \right].\end{aligned}$$

- Efron (1983, 1986) refers to the bracketed quantity as the *expected optimism* in judging the fit of a model using the same data as that which was used to construct the fit.
- In order to correct for the negative bias of  $\hat{d}(\hat{\theta}_k)$ , we must evaluate the bias adjustment represented by the expected optimism.

# Framework

- Approaches for evaluation of the bias adjustment: AIC family.
  - Obtain an asymptotic approximation for the bias adjustment.
    - AIC ( $G \in \mathcal{F}(k)$ )
    - TIC ( $G \notin \mathcal{F}(k)$ )
  - Impose structural assumptions on  $\mathcal{F}(k)$  and obtain an exact expression (or a more precise finite-sample approximation) for the bias adjustment.
    - AIC<sub>c</sub> ( $G \in \mathcal{F}(k)$ )
    - MAIC ( $G \notin \mathcal{F}(k)$ )
  - Use cross-validation or bootstrapping to estimate the bias adjustment.
    - PRESS, EIC, AIC<sub>b</sub> ( $G \notin \mathcal{F}(k)$ ) (Lecture XII)
  - Use simulation to approximate the bias adjustment.
    - AIC<sub>i</sub> ( $G \in \mathcal{F}(k)$ ) (Lecture XII)

## The Kullback Discrepancy and AIC

- Let  $f(y|\theta_k)$  denote the candidate model for  $y$ ; let  $f_i(y_i|\theta_k)$  denote the candidate model for  $y_i$ .
- Kullback Discrepancy:

$$\begin{aligned}d(G, F_{\theta_k}) &= E_G \{-2 \ln f(y|\theta_k)\} \\ &= \sum_{i=1}^n E_G \{-2 \ln f_i(y_i|\theta_k)\}\end{aligned}$$

## The Kullback Discrepancy and AIC

- Minimum discrepancy estimator:

$$\hat{\theta}_k = \operatorname{argmin}_{\theta_k \in \Theta(k)} \sum_{i=1}^n -2 \ln f_i(y_i | \theta_k).$$

- Note: The MDE is simply the MLE.

## The Kullback Discrepancy and AIC

- Overall discrepancy:

$$d(G, F_{\hat{\theta}_k}) = \sum_{i=1}^n E_G \{ -2 \ln f_i(y_i | \theta_k) \} |_{\theta_k = \hat{\theta}_k}.$$

- Expected overall discrepancy:

$$\Delta(G, k) = E_G \left\{ d(G, F_{\hat{\theta}_k}) \right\}.$$

- Estimated discrepancy:

$$\hat{d}(\hat{\theta}_k) = \sum_{i=1}^n -2 \ln f_i(y_i | \hat{\theta}_k).$$

## The Kullback Discrepancy and AIC

- Note: The estimated discrepancy

$$\sum_{i=1}^n -2 \ln f_i(y_i | \hat{\theta}_k)$$

is  $-2 \times$  the empirical log-likelihood,  $-2 \ln f(y | \hat{\theta}_k)$ , the goodness-of-fit term for AIC.

- Criteria of the AIC type:

$$-2 \ln f(y | \hat{\theta}_k) + \text{penalty term.}$$

## The SSE Discrepancy and the Gauss Criterion, $C_p$ , and FPE

- Suppose that the candidate model  $F_{\theta_k}$  corresponds to the rank  $p$  regression model

$$y_i = x_i' \beta + e_i,$$

where  $e_i \sim iid N(0, \sigma^2)$ .

- SSE Discrepancy:

$$d(G, F_{\theta_k}) = \sum_{i=1}^n E_G \{ (y_i - x_i' \beta)^2 \}.$$

## The SSE Discrepancy and the Gauss Criterion, $C_p$ , and FPE

- Minimum discrepancy estimator of  $\beta$ :

$$\hat{\beta} = \operatorname{argmin}_{\beta} \sum_{i=1}^n (y_i - x_i' \beta)^2.$$

- Note: The MDE is simply the OLS estimator.

## The SSE Discrepancy and the Gauss Criterion, $C_p$ , and FPE

- Overall discrepancy:

$$d(G, F_{\hat{\theta}_k}) = \sum_{i=1}^n E_G \left\{ (y_i - x_i' \beta)^2 \right\} \Big|_{\beta = \hat{\beta}}.$$

- Expected overall discrepancy:

$$\Delta(G, k) = E_G \left\{ d(G, F_{\hat{\theta}_k}) \right\}.$$

- Estimated discrepancy:

$$\hat{d}(\hat{\theta}_k) = \sum_{i=1}^n (y_i - x_i' \hat{\beta})^2 = \text{SSE}.$$

## The SSE Discrepancy and the Gauss Criterion, $C_p$ , and FPE

- Suppose that  $\mathcal{F}(K)$  denotes the largest candidate class in the candidate family.
- Assume that  $G \in \mathcal{F}(K)$ .
- Let  $\text{MSE}_*$  denote the mean square error for the largest candidate model.
- **Gauss criterion:**

$$G = \text{SSE} + 2p \text{MSE}_*.$$

- The Gauss criterion is exactly unbiased for  $\Delta(G, k)$ .
- The Gauss criterion yields the exact same model selections as Mallows'  $C_p$ :

$$C_p = \frac{\text{SSE}}{\text{MSE}_*} + 2p - n.$$

## The SSE Discrepancy and the Gauss Criterion, $C_p$ , and FPE

- Assume that  $G \in \mathcal{F}(k)$ .
- The following criterion is exactly unbiased for  $\Delta(G, k)$ :

$$\text{SSE} + 2p \text{MSE}.$$

- Note that we can write this criterion as follows:

$$\begin{aligned} & \text{SSE} + 2p \text{MSE} \\ &= (n + p) \text{MSE} \\ &= n \left( \frac{n + p}{n - p} \right) \hat{\sigma}^2, \end{aligned}$$

where  $\hat{\sigma}^2$  is the MLE of  $\sigma^2$ .

- This criterion is simply Akaike's FPE multiplied by  $n$ :  
 $\text{FPE} = ((n + p)/(n - p)) \hat{\sigma}^2$ .

## Simulation Study

### Study Outline:

- In each of six simulation sets, one thousand samples of size  $n$  are generated from a true regression model of the form

$$y_i = 1.0 + 4.0x_{i1} + 2.0x_{i2} + 1.0x_{i3} + 0.5x_{i4} + 0.25x_{i5} + e_i,$$

where  $e_i \sim iid N(0, 25)$ .

- For every sample, candidate models with nested design matrices of ranks  $p = 2, 3, \dots, 6$  are fit to the data. The first column of every design matrix is a vector of ones.
- The covariates are generated as *iid* replicates from a uniform  $(0, 10)$  distribution.

## Simulation Study

- The true model order is  $p_o = 6$ .
- However, because the models are comprised of effects of varying sizes, for a given sample size, the *optimal* model order is unclear.
  - For smaller sample sizes, smaller fitted models that omit the minor effects might be advantageous. When data is scarce, the estimation error associated with including minor effects might be less than the approximation error associated with excluding them.
  - For larger sample sizes, larger fitted models that include at least some of the minor effects might be advantageous. As the sample size is increased, estimation error is reduced.

## Simulation Study

- We will use the Kullback (overall) discrepancy  $d(G, F_{\hat{\theta}_k})$  as an *oracle measure* in choosing an optimal fitted model for each sample.
- We will compare the order selection patterns of AIC, MAIC, and AICc to that of the oracle.
- Additionally, we will tally the number of times that each criterion selects the same model as the oracle.
- In the six simulation sets, six different sample sizes  $n$  will be employed: 25, 50, 100, 250, 500, 1000.

## Simulation Study

Set I: Order selections with  $n = 25$ .

$p$	AIC	AIC <sub>c</sub>	MAIC	KD
2	0	1	1	8
3	33	115	152	90
4	340	493	567	288
5	350	296	233	381
6	277	95	47	233

## Simulation Study

Set II: Order selections with  $n = 50$ .

$p$	AIC	AIC <sub>c</sub>	MAIC	KD
2	0	0	0	0
3	1	2	3	6
4	171	252	285	98
5	455	480	493	357
6	373	266	219	539

## Simulation Study

Set III: Order selections with  $n = 100$ .

$p$	AIC	AIC <sub>c</sub>	MAIC	KD
2	0	0	0	0
3	0	0	0	0
4	22	24	24	9
5	439	481	490	154
6	539	495	486	837

## Simulation Study

Set IV: Order selections with  $n = 250$ .

$p$	AIC	AIC <sub>c</sub>	MAIC	KD
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0
5	126	136	140	22
6	874	864	860	978

## Simulation Study

Set V: Order selections with  $n = 500$ .

$p$	AIC	AIC <sub>c</sub>	MAIC	KD
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0
5	27	28	28	0
6	973	972	972	1000

## Simulation Study

Set VI: Order selections with  $n = 1000$ .

$p$	AIC	AIC <sub>c</sub>	MAIC	KD
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0
5	0	0	0	0
6	1000	1000	1000	1000

## Simulation Study

Agreement of criterion and oracle selections.

Set	AIC	AIC <sub>c</sub>	MAIC
I	130	115	120
II	207	149	138
III	480	436	427
IV	866	856	852
V	973	972	972
VI	1000	1000	1000