

# Econ 722 – Advanced Econometrics IV

Francis J. DiTraglia

University of Pennsylvania

# Lecture #1 – Decision Theory

Statistical Decision Theory

The James-Stein Estimator

# Decision Theoretic Preliminaries

Parameter  $\theta \in \Theta$

Unknown state of nature, from parameter space  $\Theta$

Observed Data

Observe  $X$  with distribution  $F_\theta$  from a sample space  $\mathcal{X}$

Estimator  $\hat{\theta}$

An estimator (aka a decision rule) is a function from  $\mathcal{X}$  to  $\Theta$

Loss Function  $L(\theta, \hat{\theta})$

A function from  $\Theta \times \Theta$  to  $\mathbb{R}$  that gives the cost we incur if we report  $\hat{\theta}$  when the true state of nature is  $\theta$ .

## Examples of Loss Functions

$$L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$$

squared error loss

$$L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$$

absolute error loss

$$L(\theta, \hat{\theta}) = 0 \text{ if } \theta = \hat{\theta}, 1 \text{ otherwise}$$

zero-one loss

$$L(\theta, \hat{\theta}) = \int \log \left[ \frac{f(x|\theta)}{f(x|\hat{\theta})} \right] f(x|\theta) dx$$

Kullback–Leibler loss

## (Frequentist) Risk of an Estimator $\hat{\theta}$

$$R(\theta, \hat{\theta}) = \mathbb{E}_{\theta} [L(\theta, \hat{\theta})] = \int L(\theta, \hat{\theta}(x)) dF_{\theta}(x)$$

*The frequentist decision theorist seeks to evaluate, for each  $\theta$ , how much he would “expect” to lose if he used  $\hat{\theta}(X)$  repeatedly with varying  $X$  in the problem.*

*(Berger, 1985)*

### Example: Squared Error Loss

$$R(\theta, \hat{\theta}) = \mathbb{E}_{\theta} [(\theta - \hat{\theta})^2] = \text{MSE} = \text{Var}(\hat{\theta}) + \text{Bias}_{\theta}^2(\hat{\theta})$$

# Bayes Risk and Maximum Risk

## Comparing Risk

$R(\theta, \hat{\theta})$  is a *function* of  $\theta$  rather than a single number. We want an estimator with low risk, but how can we compare?

## Maximum Risk

$$\bar{R}(\hat{\theta}) = \sup_{\theta \in \Theta} R(\theta, \hat{\theta})$$

## Bayes Risk

$$r(\pi, \hat{\theta}) = \mathbb{E}_{\pi} \left[ R(\theta, \hat{\theta}) \right], \text{ where } \pi \text{ is a prior for } \theta$$

# Bayes and Minimax Rules

Minimize the Maximum or Bayes risk over all estimators  $\tilde{\theta}$

## Minimax Rule/Estimator

$\hat{\theta}$  is **minimax** if

$$\sup_{\theta \in \Theta} R(\theta, \hat{\theta}) = \inf_{\tilde{\theta}} \sup_{\theta \in \Theta} R(\theta, \tilde{\theta})$$

## Bayes Rule/Estimator

$\hat{\theta}$  is a **Bayes rule** with respect to prior  $\pi$  if

$$r(\pi, \hat{\theta}) = \inf_{\tilde{\theta}} r(\pi, \tilde{\theta})$$

## Recall: Bayes' Theorem and Marginal Likelihood

Let  $\pi$  be a prior for  $\theta$ . By Bayes' theorem, the **posterior**  $\pi(\theta|\mathbf{x})$  is

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})}$$

where the **marginal likelihood**  $m(\mathbf{x})$  is given by

$$m(\mathbf{x}) = \int f(\mathbf{x}|\theta)\pi(\theta) d\theta$$



# Posterior Expected Loss

## Posterior Expected Loss

$$\rho(\pi(\theta|\mathbf{x}), \hat{\theta}) = \int L(\theta, \hat{\theta}) \pi(\theta|\mathbf{x}) d\theta$$

## Bayesian Decision Theory

Choose an estimator that minimizes posterior expected loss.

## Easier Calculation

Since  $m(\mathbf{x})$  does not depend on  $\theta$ , to minimize  $\rho(\pi(\theta|\mathbf{x}), \hat{\theta})$  it suffices to minimize  $\int L(\theta, \hat{\theta}) f(\mathbf{x}|\theta) \pi(\theta) d\theta$ .

## Question

Is there a relationship between Bayes risk,  $r(\pi, \hat{\theta}) \equiv \mathbb{E}_{\pi}[R(\theta, \hat{\theta})]$ , and posterior expected loss?

# Bayes Risk vs. Posterior Expected Loss

## Theorem

$$r(\pi, \hat{\theta}) = \int \rho(\pi(\theta|\mathbf{x}), \hat{\theta}(\mathbf{x})) m(\mathbf{x}) d\mathbf{x}$$

## Proof

$$\begin{aligned} r(\pi, \hat{\theta}) &= \int R(\theta, \hat{\theta}) \pi(\theta) d\theta = \int \left[ \int L(\theta, \hat{\theta}(\mathbf{x})) f(\mathbf{x}|\theta) d\mathbf{x} \right] \pi(\theta) d\theta \\ &= \int \int L(\theta, \hat{\theta}(\mathbf{x})) [f(\mathbf{x}|\theta) \pi(\theta)] d\mathbf{x} d\theta \\ &= \int \int L(\theta, \hat{\theta}(\mathbf{x})) [\pi(\theta|\mathbf{x}) m(\mathbf{x})] d\mathbf{x} d\theta \\ &= \int \left[ \int L(\theta, \hat{\theta}(\mathbf{x})) \pi(\theta|\mathbf{x}) d\theta \right] m(\mathbf{x}) d\mathbf{x} \\ &= \int \rho(\pi(\theta|\mathbf{x}), \hat{\theta}(\mathbf{x})) m(\mathbf{x}) d\mathbf{x} \end{aligned}$$

# Finding a Bayes Estimator

## Hard Problem

Find the **function**  $\hat{\theta}(\mathbf{x})$  that minimizes  $r(\pi, \hat{\theta})$ .

## Easy Problem

Find the **number**  $\hat{\theta}$  that minimizes  $\rho(\pi(\theta|\mathbf{x}), \hat{\theta})$

## Punchline

Since  $r(\pi, \hat{\theta}) = \int \rho(\pi(\theta|\mathbf{x}), \hat{\theta}(\mathbf{x})) m(\mathbf{x}) d\mathbf{x}$ , to minimize  $r(\pi, \hat{\theta})$  we can set  $\hat{\theta}(\mathbf{x})$  to be the value  $\hat{\theta}$  that minimizes  $\rho(\pi(\theta|\mathbf{x}), \hat{\theta})$ .

# Bayes Estimators for Common Loss Functions

## Zero-one Loss

For zero-one loss, the Bayes estimator is the posterior mode.

Absolute Error Loss:  $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$

For absolute error loss, the Bayes estimator is the posterior median.

Squared Error Loss:  $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$

For squared error loss, the Bayes estimator is the posterior mean.

# Derivation of Bayes Estimator for Squared Error Loss

By definition,

$$\hat{\theta} \equiv \arg \min_{a \in \Theta} \int (\theta - a)^2 \pi(\theta | \mathbf{x}) d\theta$$

Differentiating with respect to  $a$ , we have

$$\begin{aligned} 2 \int (\theta - a) \pi(\theta | \mathbf{x}) d\theta &= 0 \\ \int \theta \pi(\theta | \mathbf{x}) d\theta &= a \end{aligned}$$

## Example: Bayes Estimator for a Normal Mean

Suppose  $X \sim N(\mu, 1)$  and  $\pi$  is a  $N(a, b^2)$  prior. Then,

$$\begin{aligned}\pi(\mu|x) &\propto f(x|\mu) \times \pi(\mu) \\ &\propto \exp \left\{ -\frac{1}{2} \left[ (x - \mu)^2 + \frac{1}{b^2} (\mu - a)^2 \right] \right\} \\ &\propto \exp \left\{ -\frac{1}{2} \left[ \left( 1 + \frac{1}{b^2} \right) \mu^2 - 2 \left( x + \frac{a}{b^2} \right) \mu \right] \right\} \\ &\propto \exp \left\{ -\frac{1}{2} \left( \frac{b^2 + 1}{b^2} \right) \left[ \mu - \left( \frac{b^2 x + a}{b^2 + 1} \right) \right]^2 \right\}\end{aligned}$$

So  $\pi(\mu|x)$  is  $N(m, \omega^2)$  with  $\omega^2 = \frac{b^2}{1+b^2}$  and  $m = \omega^2 x + (1 - \omega^2)a$ .

Hence the Bayes estimator for  $\mu$  under squared error loss is

$$\hat{\theta}(X) = \frac{b^2 X + a}{1 + b^2}$$

# Minimax Analysis

## Wasserman (2004)

*The advantage of using maximum risk, despite its problems, is that it does not require one to choose a prior.*

## Berger (1986)

*Perhaps the greatest use of the minimax principle is in situations for which no prior information is available . . . but two notes of caution should be sounded. First, the minimax principle can lead to bad decision rules. . . Second, the minimax approach can be devilishly hard to implement.*

# Methods for Finding a Minimax Estimator

1. Direct Calculation
2. Guess a “Least Favorable” Prior
3. Search for an “Equalizer Rule”

Method 1 rarely applicable so focus on 2 and 3...



# The Bayes Rule for a Least Favorable Prior is Minimax

## Theorem

Let  $\hat{\theta}$  be a Bayes rule with respect to  $\pi$  and suppose that for all  $\theta \in \Theta$  we have  $R(\theta, \hat{\theta}) \leq r(\pi, \hat{\theta})$ . Then  $\hat{\theta}$  is a **minimax estimator**, and  $\pi$  is called a **least favorable prior**.

## Proof

Suppose that  $\hat{\theta}$  is not minimax. Then there exists another estimator  $\tilde{\theta}$  with  $\sup_{\theta \in \Theta} R(\theta, \tilde{\theta}) < \sup_{\theta \in \Theta} R(\theta, \hat{\theta})$ . But since

$$r(\pi, \tilde{\theta}) \equiv \mathbb{E}_{\pi} [R(\theta, \tilde{\theta})] \leq \mathbb{E}_{\pi} \left[ \sup_{\theta \in \Theta} R(\theta, \tilde{\theta}) \right] = \sup_{\theta \in \Theta} R(\theta, \tilde{\theta})$$

but this implies that  $\tilde{\theta}$  is *not* Bayes with respect to  $\pi$  since

$$r(\pi, \tilde{\theta}) \leq \sup_{\theta \in \Theta} R(\theta, \tilde{\theta}) < \sup_{\theta \in \Theta} R(\theta, \hat{\theta}) \leq r(\pi, \hat{\theta})$$

# Example of Least Favorable Prior

## Bounded Normal Mean

- ▶  $X \sim N(\theta, 1)$
- ▶ Squared error loss
- ▶  $\Theta = [-m, m]$  for  $0 < m < 1$

## Least Favorable Prior

$\pi(\theta) = 1/2$  for  $\theta \in \{-m, m\}$ , zero otherwise.

## Resulting Bayes Rule is Minimax

$$\hat{\theta}(X) = m \tanh(mX) = m \left[ \frac{\exp\{mX\} - \exp\{-mX\}}{\exp\{mX\} + \exp\{-mX\}} \right]$$

# Equalizer Rules

## Definition

An estimator  $\hat{\theta}$  is called an **equalizer rule** if its risk function is constant:  $R(\theta, \hat{\theta}) = C$  for some  $C$ .

## Theorem

If  $\hat{\theta}$  is an equalizer rule and is Bayes with respect to  $\pi$ , then  $\hat{\theta}$  is **minimax** and  $\pi$  is **least favorable**.

## Proof

$$r(\pi, \hat{\theta}) = \int R(\theta, \hat{\theta})\pi(\theta) d\theta = \int C\pi(\theta) d\theta = C$$

Hence,  $R(\theta, \hat{\theta}) \leq r(\pi, \hat{\theta})$  for all  $\theta$  so we can apply the preceding theorem.

Example:  $X_1, \dots, X_n \sim \text{iid Bernoulli}(p)$

Under a  $\text{Beta}(\alpha, \beta)$  prior with  $\alpha = \beta = \sqrt{n}/2$ ,

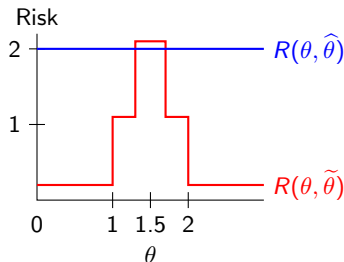
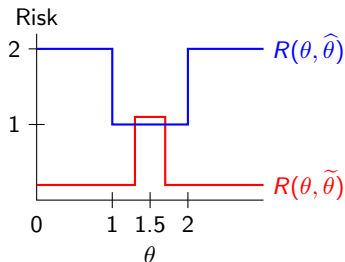
$$\hat{p}(\mathbf{x}) = \frac{n\bar{X} + \sqrt{n}/2}{n + \sqrt{n}}$$

is the Bayesian posterior mean, hence the Bayes rule under squared error loss. The risk function of  $\hat{p}$  is,

$$R(p, \hat{p}) = \frac{n}{4(n + \sqrt{n})^2}$$

which is constant in  $p$ . Hence,  $\hat{p}$  is an equalizer rule, and by the preceding theorem is minimax.

# Problems with the Minimax Principle



In the left panel,  $\tilde{\theta}$  is preferred by the minimax principle; in the right panel  $\hat{\theta}$  is preferred. But the only difference between them is that the right panel adds an additional *fixed* loss of 1 for  $1 \leq \theta \leq 2$ .

## Problems with the Minimax Principle

Suppose that  $\Theta = \{\theta_1, \theta_2\}$ ,  $\mathcal{A} = \{a_1, a_2\}$  and the loss function is:

	$a_1$	$a_2$
$\theta_1$	10	10.01
$\theta_2$	8	-8

- ▶ Minimax principle: choose  $a_1$
- ▶ Bayes: Choose  $a_2$  unless  $\pi(\theta_1) > 0.9994$

Minimax ignores the fact that under  $\theta_1$  we can never do better than a loss of 10, and tries to prevent us from incurring a tiny additional loss of 0.01

# Dominance and Admissibility

## Dominance

$\hat{\theta}$  **dominates**  $\tilde{\theta}$  with respect to  $R$  if  $R(\theta, \hat{\theta}) \leq R(\theta, \tilde{\theta})$  for all  $\theta \in \Theta$  and the inequality is strict for at least one value of  $\theta$ .

## Admissibility

$\hat{\theta}$  is **admissible** if no other estimator dominates it.

## Inadmissibility

$\hat{\theta}$  is **inadmissible** if there is an estimator that dominates it.

## Example of an Admissible Estimator

Say we want to estimate  $\theta$  from  $X \sim N(\theta, 1)$  under squared error loss. Is the estimator  $\hat{\theta}(X) = 3$  admissible?

If not, then there is a  $\tilde{\theta}$  with  $R(\theta, \tilde{\theta}) \leq R(\theta, \hat{\theta})$  for all  $\theta$ . Hence:

$$R(3, \tilde{\theta}) \leq R(3, \hat{\theta}) = \left\{ \mathbb{E} [\hat{\theta} - 3] \right\}^2 + \text{Var}(\hat{\theta}) = 0$$

Since  $R$  cannot be negative for squared error loss,

$$0 = R(3, \tilde{\theta}) = \left\{ \mathbb{E} [\tilde{\theta} - 3] \right\}^2 + \text{Var}(\tilde{\theta})$$

Therefore  $\hat{\theta} = \tilde{\theta}$ , so  $\hat{\theta}$  is admissible, although very silly!



# Bayes Rules are Admissible

## Theorem A-1

Suppose that  $\Theta$  is a discrete set and  $\pi$  gives strictly positive probability to each element of  $\Theta$ . Then, if  $\hat{\theta}$  is a Bayes rule with respect to  $\pi$ , it is admissible.

## Theorem A-2

If a Bayes rule is unique, it is admissible.

## Theorem A-3

Suppose that  $R(\theta, \hat{\theta})$  is continuous in  $\theta$  for all  $\hat{\theta}$  and that  $\pi$  gives strictly positive probability to any open subset of  $\Theta$ . Then if  $\hat{\theta}$  is a Bayes rule with respect to  $\pi$ , it is admissible.

# Admissible Equalizer Rules are Minimax

## Theorem

Let  $\hat{\theta}$  be an equalizer rule. Then if  $\hat{\theta}$  is admissible, it is minimax.

## Proof

Since  $\hat{\theta}$  is an equalizer rule,  $R(\theta, \hat{\theta}) = C$ . Suppose that  $\hat{\theta}$  is not minimax. Then there is a  $\tilde{\theta}$  such that

$$\sup_{\theta \in \Theta} R(\theta, \tilde{\theta}) < \sup_{\theta \in \Theta} R(\theta, \hat{\theta}) = C$$

But for any  $\theta$ ,  $R(\theta, \tilde{\theta}) \leq \sup_{\theta \in \Theta} R(\theta, \tilde{\theta})$ . Thus we have shown that  $\tilde{\theta}$  dominates  $\hat{\theta}$ , so that  $\hat{\theta}$  cannot be admissible.

# Minimax Implies “Nearly” Admissible

## Strong Inadmissibility

We say that  $\hat{\theta}$  is **strongly inadmissible** if there exists an estimator  $\tilde{\theta}$  and an  $\varepsilon > 0$  such that  $R(\theta, \tilde{\theta}) < R(\theta, \hat{\theta}) - \varepsilon$  for all  $\theta$ .

## Theorem

If  $\hat{\theta}$  is minimax, then it is **not** strongly inadmissible.

## Example: Sample Mean, Unbounded Parameter Space

### Theorem

Suppose that  $X_1, \dots, X_n \sim N(\theta, 1)$  with  $\Theta = \mathbb{R}$ . Under squared error loss, one can show that  $\hat{\theta} = \bar{X}$  is admissible.

### Intuition

The proof is complicated, but effectively we view this estimator as a **limit** of a of Bayes estimator with prior  $N(a, b^2)$ , as  $b^2 \rightarrow \infty$ .

### Minimaxity

Since  $R(\theta, \bar{X}) = \text{Var}(\bar{X}) = 1/n$ , we see that  $\bar{X}$  is an equalizer rule. Since it is admissible, it is therefore minimax.

# Recall: Gauss-Markov Theorem

## Linear Regression Model

$$\mathbf{y} = X\beta + \epsilon, \quad \mathbb{E}[\epsilon|X] = \mathbf{0}$$

## Best Linear Unbiased Estimator

- ▶  $\text{Var}(\epsilon|X) = \sigma^2 I \Rightarrow$  then OLS has lowest variance among linear, unbiased estimators of  $\beta$ .
- ▶  $\text{Var}(\epsilon|X) \neq \sigma^2 I \Rightarrow$  then GLS gives a lower variance estimator.

What if we consider biased estimators and squared error loss?

## Multiple Normal Means: $X \sim N(\theta, I)$

### Goal

Estimate the  $p$ -vector  $\theta$  using  $X$  with  $L(\theta, \hat{\theta}) = \|\hat{\theta} - \theta\|^2$ .

### Maximum Likelihood Estimator $\hat{\theta}$

MLE = sample mean, but only one observation:  $\hat{\theta} = X$ .

### Risk of $\hat{\theta}$

$$(\hat{\theta} - \theta)' (\hat{\theta} - \theta) = (X - \theta)' (X - \theta) = \sum_{i=1}^p (X_i - \theta_i)^2 \sim \chi_p^2$$

Since  $\mathbb{E}[\chi_p^2] = p$ , we have  $R(\theta, \hat{\theta}) = p$ .

## Multiple Normal Means: $X \sim N(\theta, I)$

### James-Stein Estimator

$$\hat{\theta}^{JS} = \hat{\theta} \left( 1 - \frac{p-2}{\hat{\theta}'\hat{\theta}} \right) = X - \frac{(p-2)X}{X'X}$$

- ▶ Shrinks components of sample mean vector towards zero
- ▶ More elements in  $\theta \Rightarrow$  more shrinkage
- ▶ MLE close to zero ( $\hat{\theta}'\hat{\theta}$  small) gives more shrinkage

## MSE of James-Stein Estimator

$$\begin{aligned}R(\theta, \hat{\theta}^{JS}) &= \mathbb{E} \left[ (\hat{\theta}^{JS} - \theta)' (\hat{\theta}^{JS} - \theta) \right] \\&= \mathbb{E} \left[ \left\{ (X - \theta) - \frac{(p-2)X}{X'X} \right\}' \left\{ (X - \theta) - \frac{(p-2)X}{X'X} \right\} \right] \\&= \mathbb{E} [(X - \theta)' (X - \theta)] - 2(p-2) \mathbb{E} \left[ \frac{X'(X - \theta)}{X'X} \right] \\&\quad + (p-2)^2 \mathbb{E} \left[ \frac{1}{X'X} \right] \\&= p - 2(p-2) \mathbb{E} \left[ \frac{X'(X - \theta)}{X'X} \right] + (p-2)^2 \mathbb{E} \left[ \frac{1}{X'X} \right]\end{aligned}$$

Using fact that  $R(\theta, \hat{\theta}) = p$



# Simplifying the Second Term

## Writing Numerator as a Sum

$$\mathbb{E} \left[ \frac{X'(X - \theta)}{X'X} \right] = \mathbb{E} \left[ \frac{\sum_{i=1}^p X_i (X_i - \theta_i)}{X'X} \right] = \sum_{i=1}^p \mathbb{E} \left[ \frac{X_i (X_i - \theta_i)}{X'X} \right]$$

For  $i = 1, \dots, p$

$$\mathbb{E} \left[ \frac{X_i (X_i - \theta_i)}{X'X} \right] = \mathbb{E} \left[ \frac{X'X - 2X_i^2}{(X'X)^2} \right]$$

Not obvious: integration by parts, expectation as a  $p$ -fold integral,  $X \sim N(\theta, I)$

## Combining

$$\begin{aligned} \mathbb{E} \left[ \frac{X'(X - \theta)}{X'X} \right] &= \sum_{i=1}^p \mathbb{E} \left[ \frac{X'X - 2X_i^2}{(X'X)^2} \right] = p \mathbb{E} \left[ \frac{1}{X'X} \right] - 2 \mathbb{E} \left[ \frac{\sum_{i=1}^p X_i^2}{(X'X)^2} \right] \\ &= p \mathbb{E} \left[ \frac{1}{X'X} \right] - 2 \mathbb{E} \left[ \frac{X'X}{(X'X)^2} \right] = (p - 2) \mathbb{E} \left[ \frac{1}{X'X} \right] \end{aligned}$$

## The MLE is Inadmissible when $p \geq 3$

$$\begin{aligned} R\left(\theta, \hat{\theta}^{JS}\right) &= p - 2(p-2) \left\{ (p-2) \mathbb{E} \left[ \frac{1}{X'X} \right] \right\} + (p-2)^2 \mathbb{E} \left[ \frac{1}{X'X} \right] \\ &= p - (p-2)^2 \mathbb{E} \left[ \frac{1}{X'X} \right] \end{aligned}$$

- ▶  $\mathbb{E}[1/(X'X)]$  exists and is positive whenever  $p \geq 3$
- ▶  $(p-2)^2$  is always positive
- ▶ Hence, second term in the MSE expression is *negative*
- ▶ First term is MSE of the MLE

Therefore James-Stein strictly dominates MLE whenever  $p \geq 3$ !

## James-Stein More Generally

- ▶ Our example was specific, but the result is general:
  - ▶ MLE is inadmissible under quadratic loss in regression model with at least three regressors.
  - ▶ Note, however, that this is MSE for the *full parameter vector*
- ▶ James-Stein estimator is also inadmissible!
  - ▶ Dominated by “positive-part” James-Stein estimator:

$$\hat{\beta}^{JS} = \hat{\beta} \left[ 1 - \frac{(p-2)\hat{\sigma}^2}{\hat{\beta}'X'X\hat{\beta}} \right]_+$$

- ▶  $\hat{\beta} = \text{OLS}$ ,  $(x)_+ = \max(x, 0)$ ,  $\hat{\sigma}^2 = \text{usual OLS-based estimator}$
- ▶ Stops us from shrinking *past* zero to get a negative estimate for an element of  $\beta$  with a small OLS estimate.
- ▶ Positive-part James-Stein isn't admissible either!

# Lecture #2 – Model Selection I

Kullback-Leibler Divergence

Bias of Maximized Sample Log-Likelihood

Review of Asymptotics for Mis-specified MLE

Deriving AIC and TIC

Corrected AIC ( $AIC_c$ )

Mallow's  $C_p$

# Kullback-Leibler (KL) Divergence

## Motivation

How well does a given density  $f(y)$  approximate an unknown true density  $g(y)$ ? Use this to select between parametric models.

## Definition

$$\text{KL}(g; f) = \underbrace{\mathbb{E}_G \left[ \log \left\{ \frac{g(Y)}{f(Y)} \right\} \right]}_{\text{True density on top}} = \underbrace{\mathbb{E}_G [\log g(Y)]}_{\substack{\text{Depends only on truth} \\ \text{Fixed across models}}} - \underbrace{\mathbb{E}_G [\log f(Y)]}_{\text{Expected log-likelihood}}$$

## Properties

- ▶ Not symmetric:  $\text{KL}(g; f) \neq \text{KL}(f; g)$
- ▶ By Jensen's Inequality:  $\text{KL}(g; f) \geq 0$  (strict iff  $g = f$  a.e.)
- ▶ Minimize KL  $\iff$  Maximize Expected log-likelihood

# KL Divergence and Mis-specified MLE

Pseudo-true Parameter Value  $\theta_0$

$$\hat{\theta}_{MLE} \xrightarrow{P} \theta_0 \equiv \arg \min_{\theta \in \Theta} \text{KL}(g; f_{\theta}) = \arg \max_{\theta \in \Theta} \mathbb{E}_G[\log f(Y|\theta)]$$

What if  $f_{\theta}$  is correctly specified?

If  $g = f_{\theta}$  for some  $\theta$  then  $\text{KL}(g; f_{\theta})$  is minimized at zero.

Goal: Compare Mis-specified Models

$$\mathbb{E}_G [\log f(Y|\theta_0)] \quad \text{versus} \quad \mathbb{E}_G [\log h(Y|\gamma_0)]$$

where  $\theta_0$  is the pseudo-true parameter value for  $f_{\theta}$  and  $\gamma_0$  is the pseudo-true parameter value for  $h_{\gamma}$ .

# How to Estimate Expected Log Likelihood?

For simplicity:  $Y_1, \dots, Y_n \sim \text{iid } g(y)$

## Unbiased but Infeasible

$$\mathbb{E}_G \left[ \frac{1}{T} \ell(\theta_0) \right] = \mathbb{E}_G \left[ \frac{1}{T} \sum_{t=1}^T \log f(Y_t | \theta_0) \right] = \mathbb{E}_G [\log f(Y | \theta_0)]$$

## Biased but Feasible

$T^{-1} \ell(\hat{\theta}_{MLE})$  is a **biased** estimator of  $\mathbb{E}_G[\log f(Y | \theta_0)]$ .

## Intuition for the Bias

$T^{-1} \ell(\hat{\theta}_{MLE}) > T^{-1} \ell(\theta_0)$  unless  $\hat{\theta}_{MLE} = \theta_0$ . Maximized sample log-like. is an **overly optimistic** estimator of expected log-like.

# What to do about this bias?

1. General-purpose asymptotic approximation of “degree of over-optimism” of maximized sample log-likelihood.
  - ▶ Takeuchi's Information Criterion (TIC)
  - ▶ Akaike's Information Criterion (AIC)
2. Problem-specific finite sample approach, assuming  $g \in f_\theta$ .
  - ▶ Corrected AIC ( $AIC_c$ ) of Hurvich and Tsai (1989)

## Tradeoffs

TIC is most general and makes weakest assumptions, but requires very large  $T$  to work well. AIC is a good approximation to TIC that requires less data. Both AIC and TIC perform poorly when  $T$  is small relative to the number of parameters, hence  $AIC_c$ .



# Recall: Asymptotics for Mis-specified ML Estimation

Model  $f(y|\theta)$ , pseudo-true parameter  $\theta_0$ . For simplicity  $Y_1, \dots, Y_T \sim \text{iid } g(y)$ .

## Fundamental Expansion

$$\sqrt{T}(\hat{\theta} - \theta_0) = J^{-1} \left( \sqrt{T} \bar{U}_T \right) + o_p(1)$$

$$J = -\mathbb{E}_G \left[ \frac{\partial \log f(Y|\theta_0)}{\partial \theta \partial \theta'} \right], \quad \bar{U}_T = \frac{1}{T} \sum_{t=1}^T \frac{\partial \log f(Y_t|\theta_0)}{\partial \theta}$$

## Central Limit Theorem

$$\sqrt{T} \bar{U}_T \rightarrow_d U \sim N_p(0, K), \quad K = \text{Var}_G \left[ \frac{\partial \log f(Y|\theta_0)}{\partial \theta} \right]$$

$$\sqrt{T}(\hat{\theta} - \theta_0) \rightarrow_d J^{-1} U \sim N_p(0, J^{-1} K J^{-1})$$

## Information Matrix Equality

If  $g = f_\theta$  for some  $\theta \in \Theta$  then  $K = J \implies \text{AVAR}(\hat{\theta}) = J^{-1}$

# Bias Relative to Infeasible Plug-in Estimator

## Definition of Bias Term $B$

$$B = \underbrace{\frac{1}{T} \ell(\hat{\theta})}_{\text{feasible over-optimistic}} - \underbrace{\int g(y) \log f(y|\hat{\theta}) dy}_{\text{uses data only once infeas. not over-optimistic}}$$

## Question to Answer

On average, over the sampling distribution of  $\hat{\theta}$ , how large is  $B$ ?

AIC and TIC construct an asymptotic approximation of  $\mathbb{E}[B]$ .

# Derivation of AIC/TIC

## Step 1: Taylor Expansion

$$B = \bar{Z}_T + (\hat{\theta} - \theta_0)' J(\hat{\theta} - \theta_0) + o_p(T^{-1})$$

$$\bar{Z}_T = \frac{1}{T} \sum_{t=1}^T \{\log f(Y_t|\theta_0) - \mathbb{E}_G[\log f(Y|\theta_0)]\}$$

## Step 2: $\mathbb{E}[\bar{Z}_T] = 0$

$$\mathbb{E}[B] \approx \mathbb{E} \left[ (\hat{\theta} - \theta_0)' J(\hat{\theta} - \theta_0) \right]$$

## Step 3: $\sqrt{T}(\hat{\theta} - \theta_0) \rightarrow_d J^{-1}U$

$$T(\hat{\theta} - \theta_0)' J(\hat{\theta} - \theta_0) \rightarrow_d U' J^{-1}U$$

## Derivation of AIC/TIC Continued...

Step 3:  $\sqrt{T}(\hat{\theta} - \theta_0) \rightarrow_d J^{-1}U$

$$T(\hat{\theta} - \theta_0)'J(\hat{\theta} - \theta_0) \rightarrow_d U'J^{-1}U$$

Step 4:  $U \sim N_p(0, K)$

$$\mathbb{E}[B] \approx \frac{1}{T}\mathbb{E}[U'J^{-1}U] = \frac{1}{T}\text{tr}\{J^{-1}K\}$$

Final Result:

$T^{-1}\text{tr}\{J^{-1}K\}$  is an asymp. unbiased estimator of the over-optimism of  $T^{-1}\ell(\hat{\theta})$  relative to  $\int g(y) \log f(y|\hat{\theta}) dy$ .

# TIC and AIC

## Takeuchi's Information Criterion

Multiply by  $2T$ , estimate  $J, K \Rightarrow \text{TIC} = 2 \left[ \ell(\hat{\theta}) - \text{tr} \left\{ \hat{J}^{-1} \hat{K} \right\} \right]$

## Akaike's Information Criterion

If  $g = f_{\theta}$  then  $J = K \Rightarrow \text{tr} \{ J^{-1} K \} = p \Rightarrow \text{AIC} = 2 \left[ \ell(\hat{\theta}) - p \right]$

## Contrasting AIC and TIC

Technically, AIC requires that all models under consideration are at least correctly specified while TIC doesn't. But  $J^{-1}K$  is hard to estimate, and if a model is badly mis-specified,  $\ell(\hat{\theta})$  dominates.

## Corrected AIC ( $AIC_c$ ) – Hurvich & Tsai (1989)

### Idea Behind $AIC_c$

Asymptotic approximation used for AIC/TIC works poorly if  $p$  is too large relative to  $T$ . Try exact, finite-sample approach instead.

Assumption: True DGP

$$\mathbf{y} = \mathbf{X}\beta_0 + \varepsilon, \quad \varepsilon \sim N(\mathbf{0}, \sigma_0^2 \mathbf{I}_T), \quad k \text{ Regressors}$$

Can Show That

$$KL(g, f) = \frac{T}{2} \left[ \frac{\sigma_0^2}{\sigma_1^2} - \log \left( \frac{\sigma_0^2}{\sigma_1^2} \right) - 1 \right] + \left( \frac{1}{2\sigma_1^2} \right) (\beta_0 - \beta_1)' \mathbf{X}' \mathbf{X} (\beta_0 - \beta_1)$$

Where  $f$  is a normal regression model with parameters  $(\beta_1, \sigma_1^2)$  that might not be the true parameters.

## But how can we use this?

$$KL(g, f) = \frac{T}{2} \left[ \frac{\sigma_0^2}{\sigma_1^2} - \log \left( \frac{\sigma_0^2}{\sigma_1^2} \right) - 1 \right] + \left( \frac{1}{2\sigma_1^2} \right) (\beta_0 - \beta_1)' \mathbf{X}' \mathbf{X} (\beta_0 - \beta_1)$$

1. Would need to know  $(\beta_1, \sigma_1^2)$  for **candidate model**.
  - ▶ Easy: just use MLE  $(\hat{\beta}_1, \hat{\sigma}_1^2)$
2. Would need to know  $(\beta_0, \sigma_0^2)$  for **true model**.
  - ▶ Very hard! The whole problem is that we don't know these!

### Hurvich & Tsai (1989) Assume:

- ▶ Every candidate model is **at least correctly specified**
- ▶ Implies any candidate estimator  $(\hat{\beta}, \hat{\sigma}^2)$  is consistent for truth.

## Deriving the Corrected AIC

Since  $(\hat{\beta}, \hat{\sigma}^2)$  are random, look at  $\mathbb{E}[\widehat{KL}]$ , where

$$\widehat{KL} = \frac{T}{2} \left[ \frac{\sigma_0^2}{\hat{\sigma}^2} - \log \left( \frac{\sigma_0^2}{\hat{\sigma}^2} \right) - 1 \right] + \left( \frac{1}{2\hat{\sigma}^2} \right) (\hat{\beta} - \beta_0)' \mathbf{X}' \mathbf{X} (\hat{\beta} - \beta_0)$$

Finite-sample theory for correctly spec. normal regression model:

$$\mathbb{E}[\widehat{KL}] = \frac{T}{2} \left\{ \frac{T+k}{T-k-2} - \log(\sigma_0^2) + \mathbb{E}[\log \hat{\sigma}^2] - 1 \right\}$$

Eliminate constants and scaling, unbiased estimator of  $\mathbb{E}[\log \hat{\sigma}^2]$ :

$$\text{AIC}_c = \log \hat{\sigma}^2 + \frac{T+k}{T-k-2}$$

a finite-sample unbiased estimator of KL for model comparison



## Motivation: Predict $\mathbf{y}$ from $\mathbf{x}$ via Linear Regression

$$\underset{(T \times 1)}{\mathbf{y}} = \underset{(T \times K)}{\mathbf{X}} \underset{(K \times 1)}{\boldsymbol{\beta}} + \boldsymbol{\epsilon}$$

$$\mathbb{E}[\boldsymbol{\epsilon}|\mathbf{X}] = 0, \quad \text{Var}(\boldsymbol{\epsilon}|\mathbf{X}) = \sigma^2 \mathbf{I}$$

- ▶ If  $\boldsymbol{\beta}$  were known, could never achieve lower MSE than by using all regressors to predict.
- ▶ But  $\boldsymbol{\beta}$  is unknown so we have to estimate it from data  $\Rightarrow$  bias-variance tradeoff.
- ▶ Could make sense to exclude regressors with small coefficients: add small bias but reduce variance.

# Operationalizing the Bias-Variance Tradeoff Idea

## Mallow's $C_p$

Approximate the predictive MSE of each model relative to the infeasible optimum in which  $\beta$  is known.

## Notation

- ▶ Model index  $m$  and regressor matrix  $\mathbf{X}_m$
- ▶ Corresponding OLS estimator  $\hat{\beta}_m$  padded out with zeros
- ▶  $\mathbf{X}\hat{\beta}_m = \mathbf{X}_{(-m)}\mathbf{0} + \mathbf{X}_m [(\mathbf{X}'_m \mathbf{X}_m)^{-1} \mathbf{X}'_m \mathbf{y}] = \mathbf{P}_m \mathbf{y}$

# In-sample versus Out-of-sample Prediction Error

Why not compare  $RSS(m)$ ?

In-sample prediction error:  $RSS(m) = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_m)'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_m)$

From your Problem Set

RSS cannot decrease even if we add irrelevant regressors. Thus in-sample prediction error is an **overly optimistic** estimate of out-of-sample prediction error.

Bias-Variance Tradeoff

Out-of-sample performance of full model (using all regressors) could be very poor if there is a lot of estimation uncertainty associated with regressors that aren't very predictive.

# Predictive MSE of $\mathbf{X}\hat{\boldsymbol{\beta}}_m$ relative to infeasible optimum $\mathbf{X}\boldsymbol{\beta}$

Step 1: Algebra

$$\begin{aligned}\mathbf{X}\hat{\boldsymbol{\beta}}_m - \mathbf{X}\boldsymbol{\beta} &= \mathbf{P}_m\mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \mathbf{P}_m(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - (\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{P}_m\boldsymbol{\epsilon} - (\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta}\end{aligned}$$

Step 2:  $\mathbf{P}_m$  and  $(\mathbf{I} - \mathbf{P}_m)$  are both symmetric and idempotent, and orthogonal to each other

$$\begin{aligned}\left\|\mathbf{X}\hat{\boldsymbol{\beta}}_m - \mathbf{X}\boldsymbol{\beta}\right\|^2 &= \{\mathbf{P}_m\boldsymbol{\epsilon} - (\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta}\}' \{\mathbf{P}_m\boldsymbol{\epsilon} + (\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta}\} \\ &= \boldsymbol{\epsilon}'\mathbf{P}_m'\mathbf{P}_m\boldsymbol{\epsilon} - \boldsymbol{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{P}_m)'\mathbf{P}_m\boldsymbol{\epsilon} - \boldsymbol{\epsilon}'\mathbf{P}_m'(\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta} \\ &\quad + \boldsymbol{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{P}_m)(\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\epsilon}'\mathbf{P}_m\boldsymbol{\epsilon} + \boldsymbol{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{P}_m)\mathbf{X}\boldsymbol{\beta}\end{aligned}$$

## Predictive MSE of $\mathbf{X}\hat{\boldsymbol{\beta}}_m$ relative to infeasible optimum $\mathbf{X}\boldsymbol{\beta}$

Step 3: Expectation of Step 2 conditional on  $\mathbf{X}$

$$\begin{aligned}\text{MSE}(m|\mathbf{X}) &= \mathbb{E} \left[ (\mathbf{X}\hat{\boldsymbol{\beta}}_m - \mathbf{X}\boldsymbol{\beta})' (\mathbf{X}\hat{\boldsymbol{\beta}}_m - \mathbf{X}\boldsymbol{\beta}) | \mathbf{X} \right] \\ &= \mathbb{E} [\boldsymbol{\epsilon}' \mathbf{P}_m \boldsymbol{\epsilon} | \mathbf{X}] + \mathbb{E} [\boldsymbol{\beta}' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \boldsymbol{\beta} | \mathbf{X}] \\ &= \mathbb{E} [\text{tr} \{ \boldsymbol{\epsilon}' \mathbf{P}_m \boldsymbol{\epsilon} \} | \mathbf{X}] + \boldsymbol{\beta}' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \boldsymbol{\beta} \\ &= \text{tr} \{ \mathbb{E} [\boldsymbol{\epsilon} \boldsymbol{\epsilon}' | \mathbf{X}] \mathbf{P}_m \} + \boldsymbol{\beta}' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \boldsymbol{\beta} \\ &= \text{tr} \{ \sigma^2 \mathbf{P}_m \} + \boldsymbol{\beta}' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \boldsymbol{\beta} \\ &= \sigma^2 k_m + \boldsymbol{\beta}' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \boldsymbol{\beta}\end{aligned}$$

where  $k_m$  denotes the number of regressors in  $\mathbf{X}_m$  and

$$\text{tr}(\mathbf{P}_m) = \text{tr} \left\{ \mathbf{X}_m (\mathbf{X}_m' \mathbf{X}_m)^{-1} \mathbf{X}_m' \right\} = \text{tr} \left\{ \mathbf{X}_m' \mathbf{X}_m (\mathbf{X}_m' \mathbf{X}_m)^{-1} \right\} = \text{tr}(\mathbf{I}_m)$$

Now we know the MSE of a given model...

$$\text{MSE}(m|\mathbf{X}) = \sigma^2 k_m + \beta' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \beta$$

### Bias-Variance Tradeoff

- ▶ Smaller Model  $\Rightarrow \sigma^2 k_m$  smaller: less estimation uncertainty.
- ▶ Bigger Model  $\Rightarrow \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} = \|(\mathbf{I} - \mathbf{P}_m) \mathbf{X}\|^2$  is in general smaller: less (squared) bias.

### Mallow's $C_p$

- ▶ Problem: MSE formula is infeasible since it involves  $\beta$  and  $\sigma^2$ .
- ▶ Solution: Mallow's  $C_p$  constructs an unbiased estimator.
- ▶ Idea: what about plugging in  $\hat{\beta}$  to estimate second term?

## What if we plug in $\hat{\beta}$ to estimate the second term?

For the missing algebra in Step 4, see the lecture notes.

### Notation

Let  $\hat{\beta}$  denote the full model estimator and  $\mathbf{P}$  be the corresponding projection matrix:  $\mathbf{X}\hat{\beta} = \mathbf{P}\mathbf{y}$ .

### Crucial Fact

$\text{span}(\mathbf{X}_m)$  is a subspace of  $\text{span}(\mathbf{X})$ , so  $\mathbf{P}_m\mathbf{P} = \mathbf{P}\mathbf{P}_m = \mathbf{P}_m$ .

### Step 4: Algebra using the preceding fact

$$\mathbb{E} \left[ \hat{\beta}' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \hat{\beta} | \mathbf{X} \right] = \dots = \beta' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \beta + \mathbb{E} \left[ \epsilon' (\mathbf{P} - \mathbf{P}_m) \epsilon | \mathbf{X} \right]$$

## Substituting $\hat{\beta}$ doesn't work...

Step 5: Use “Trace Trick” on second term from Step 4

$$\begin{aligned}\mathbb{E}[\epsilon'(\mathbf{P} - \mathbf{P}_m)\epsilon|\mathbf{X}] &= \mathbb{E}[\text{tr}\{\epsilon'(\mathbf{P} - \mathbf{P}_m)\epsilon\}|\mathbf{X}] \\ &= \text{tr}\{\mathbb{E}[\epsilon\epsilon'|\mathbf{X}](\mathbf{P} - \mathbf{P}_m)\} \\ &= \text{tr}\{\sigma^2(\mathbf{P} - \mathbf{P}_m)\} \\ &= \sigma^2(\text{trace}\{\mathbf{P}\} - \text{trace}\{\mathbf{P}_m\}) \\ &= \sigma^2(K - k_m)\end{aligned}$$

where  $K$  is the total number of regressors in  $\mathbf{X}$

### Bias of Plug-in Estimator

$$\mathbb{E}\left[\hat{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{P}_m)\mathbf{X}\hat{\beta}|\mathbf{X}\right] = \underbrace{\beta'\mathbf{X}'(\mathbf{I} - \mathbf{P}_m)\mathbf{X}\beta}_{\text{Truth}} + \underbrace{\sigma^2(K - k_m)}_{\text{Bias}}$$



## Putting Everything Together: Mallows's $C_p$

Want An Unbiased Estimator of This:

$$\text{MSE}(m|\mathbf{X}) = \sigma^2 k_m + \boldsymbol{\beta}' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \boldsymbol{\beta}$$

Previous Slide:

$$\mathbb{E} \left[ \hat{\boldsymbol{\beta}}' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \hat{\boldsymbol{\beta}} | \mathbf{X} \right] = \boldsymbol{\beta}' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \boldsymbol{\beta} + \sigma^2 (K - k_m)$$

End Result:

$$\begin{aligned} \text{MC}(m) &= \hat{\sigma}^2 k_m + \left[ \hat{\boldsymbol{\beta}}' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \hat{\boldsymbol{\beta}} - \hat{\sigma}^2 (K - k_m) \right] \\ &= \hat{\boldsymbol{\beta}}' \mathbf{X}' (\mathbf{I} - \mathbf{P}_m) \mathbf{X} \hat{\boldsymbol{\beta}} + \hat{\sigma}^2 (2k_m - K) \end{aligned}$$

is an unbiased estimator of MSE, with  $\hat{\sigma}^2 = \mathbf{y}'(\mathbf{I} - \mathbf{P})\mathbf{y}/(T - K)$

## Why is this different from the textbook formula?

Just algebra, but tedious. . .

$$\begin{aligned}\text{MC}(m) - 2\hat{\sigma}^2 k_m &= \hat{\beta}' X' (\mathbf{I} - P_M) X \hat{\beta} - K \hat{\sigma}^2 \\ &\vdots \\ &= \mathbf{y}' (\mathbf{I} - P_M) \mathbf{y} - T \hat{\sigma}^2 \\ &= \text{RSS}(m) - T \hat{\sigma}^2\end{aligned}$$

Therefore:

$$\text{MC}(m) = \text{RSS}(m) + \hat{\sigma}^2(2k_m - T)$$

Divide Through by  $\hat{\sigma}^2$ :

$$C_p(m) = \frac{\text{RSS}(m)}{\hat{\sigma}^2} + 2k_m - T$$

Tells us how to adjust RSS for number of regressors. . .

# Lecture #3 – Model Selection II

Bayesian Model Comparison

Bayesian Information Criterion (BIC)

K-fold Cross-validation

Asymptotic Equivalence Between LOO-CV and TIC

# Bayesian Model Comparison: Marginal Likelihoods

## Bayes' Rule for Model $m \in \mathcal{M}$

$$\underbrace{\pi(\boldsymbol{\theta}|\mathbf{y}, m)}_{\text{Posterior}} \propto \underbrace{\pi(\boldsymbol{\theta}|m)}_{\text{Prior}} \underbrace{f(\mathbf{y}|\boldsymbol{\theta}, m)}_{\text{Likelihood}}$$
$$\underbrace{f(\mathbf{y}|m)}_{\text{Marginal Likelihood}} = \int_{\Theta} \pi(\boldsymbol{\theta}|m) f(\mathbf{y}|\boldsymbol{\theta}, m) \, d\boldsymbol{\theta}$$

## Posterior Model Probability for $m \in \mathcal{M}$

$$P(m|\mathbf{y}) = \frac{P(m)f(\mathbf{y}|m)}{f(\mathbf{y})} = \frac{\int_{\Theta} P(m)f(\mathbf{y}, \boldsymbol{\theta}|m) \, d\boldsymbol{\theta}}{f(\mathbf{y})} = \frac{P(m)}{f(\mathbf{y})} \int_{\Theta} \pi(\boldsymbol{\theta}|m)f(\mathbf{y}|\boldsymbol{\theta}, m) \, d\boldsymbol{\theta}$$

where  $P(m)$  is the **prior model probability** and  $f(\mathbf{y})$  is constant across models.

# Laplace (aka Saddlepoint) Approximation

Suppress model index  $m$  for simplicity.

General Case: for  $T$  large...

$$\int_{\Theta} g(\boldsymbol{\theta}) \exp\{T \cdot h(\boldsymbol{\theta})\} d\boldsymbol{\theta} \approx \left(\frac{2\pi}{T}\right)^{p/2} \exp\{T \cdot h(\boldsymbol{\theta}_0)\} g(\boldsymbol{\theta}_0) |H(\boldsymbol{\theta}_0)|^{-1/2}$$

$$p = \dim(\boldsymbol{\theta}), \quad \boldsymbol{\theta}_0 = \arg \max_{\boldsymbol{\theta} \in \Theta} h(\boldsymbol{\theta}), \quad H(\boldsymbol{\theta}_0) = -\frac{\partial^2 h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$$

Use to Approximate Marginal Likelihood

$$h(\boldsymbol{\theta}) = \frac{\ell(\boldsymbol{\theta})}{T} = \frac{1}{T} \sum_{t=1}^T \log f(Y_t | \boldsymbol{\theta}), \quad H(\boldsymbol{\theta}) = J_T(\boldsymbol{\theta}) = -\frac{1}{T} \sum_{t=1}^T \frac{\partial^2 \log f(Y_t | \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}, \quad g(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$$

and substitute  $\hat{\boldsymbol{\theta}}_{MLE}$  for  $\boldsymbol{\theta}_0$

# Laplace Approximation to Marginal Likelihood

Suppress model index  $m$  for simplicity.

$$\int_{\Theta} \pi(\boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta}) \, d\boldsymbol{\theta} \approx \left(\frac{2\pi}{T}\right)^{p/2} \exp\left\{\ell(\hat{\boldsymbol{\theta}}_{MLE})\right\} \pi(\hat{\boldsymbol{\theta}}_{MLE}) \left|J_T(\hat{\boldsymbol{\theta}}_{MLE})\right|^{-1/2}$$

$$\ell(\boldsymbol{\theta}) = \sum_{t=1}^T \log f(Y_t|\boldsymbol{\theta}), \quad H(\boldsymbol{\theta}) = J_T(\boldsymbol{\theta}) = -\frac{1}{T} \sum_{t=1}^T \frac{\partial^2 \log f(Y_t|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$$

# Bayesian Information Criterion

$$\int_{\Theta} \pi(\boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta}) \, d\boldsymbol{\theta} \approx \left(\frac{2\pi}{T}\right)^{p/2} \exp\left\{\ell(\hat{\boldsymbol{\theta}}_{MLE})\right\} \pi(\hat{\boldsymbol{\theta}}_{MLE}) \left|J_T(\hat{\boldsymbol{\theta}}_{MLE})\right|^{-1/2}$$

Take Logs and Multiply by 2

$$2 \log f(\mathbf{y}|\boldsymbol{\theta}) \approx \underbrace{2\ell(\hat{\boldsymbol{\theta}}_{MLE})}_{O_p(T)} - \underbrace{p \log(T)}_{O(\log T)} + \underbrace{p \log(2\pi) + \log \pi(\hat{\boldsymbol{\theta}}) - \log |J_T(\hat{\boldsymbol{\theta}})|}_{O_p(1)}$$

The BIC

Assume uniform prior over **models** and ignore lower order terms:

$$\text{BIC}(m) = 2 \log f(\mathbf{y}|\hat{\boldsymbol{\theta}}, m) - p_m \log(T)$$

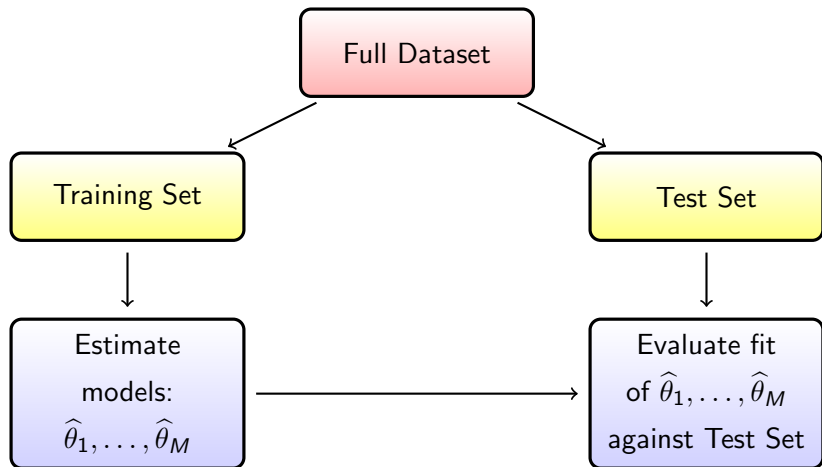
large-sample Frequentist approx. to Bayesian marginal likelihood

# Model Selection using a Hold-out Sample

- ▶ The real problem is **double** use of the data: first for estimation, then for model comparison.
  - ▶ Maximized sample log-likelihood is an overly optimistic estimate of expected log-likelihood and hence KL-divergence
  - ▶ In-sample squared prediction error is an overly optimistic estimator of out-of-sample squared prediction error
- ▶ AIC/TIC,  $AIC_c$ , BIC,  $C_p$  **penalize** sample log-likelihood or RSS to compensate.
- ▶ Another idea: **don't re-use the same data!**

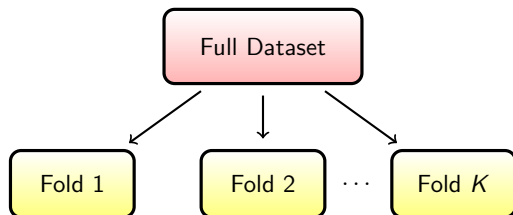


## Hold-out Sample: Partition the Full Dataset



Unfortunately this is extremely wasteful of data...

## K-fold Cross-Validation: “Pseudo-out-of-sample”



### Step 1

Randomly partition full dataset into  $K$  folds of approx. equal size.

### Step 2

Treat  $k^{\text{th}}$  fold as a hold-out sample and estimate model using all observations **except** those in fold  $k$ : yielding estimator  $\hat{\theta}(-k)$ .

# $K$ -fold Cross-Validation: “Pseudo-out-of-sample”

## Step 2

Treat  $k^{\text{th}}$  fold as a hold-out sample and estimate model using all observations **except** those in fold  $k$ : yielding estimator  $\hat{\theta}(-k)$ .

## Step 3

Repeat Step 2 for each  $k = 1, \dots, K$ .

## Step 4

For each  $t$  calculate the prediction  $\hat{y}_t^{-k(t)}$  of  $y_t$  based on  $\hat{\theta}(-k(t))$ , the estimator that excluded observation  $t$ .

## $K$ -fold Cross-Validation: “Pseudo-out-of-sample”

### Step 4

For each  $t$  calculate the prediction  $\hat{y}_t^{-k(t)}$  of  $y_t$  based on  $\hat{\theta}(-k(t))$ , the estimator that excluded observation  $t$ .

### Step 5

Define  $CV_K = \frac{1}{T} \sum_{t=1}^T L(y_t, \hat{y}_t^{-k(t)})$  where  $L$  is a loss function.

### Step 5

Repeat for each model & choose  $m$  to minimize  $CV_K(m)$ .

CV uses each observation for parameter estimation and model evaluation but never at the same time!

# Cross-Validation (CV): Some Details

## Which Loss Function?

- ▶ For regression squared error loss makes sense
- ▶ For classification (discrete prediction) could use zero-one loss.
- ▶ Can also use log-likelihood/KL-divergence as a loss function. . .

## How Many Folds?

- ▶ One extreme:  $K = 2$ . Closest to Training/Test idea.
- ▶ Other extreme:  $K = T$  **Leave-one-out** CV (LOO-CV).
- ▶ Computationally expensive model  $\Rightarrow$  may prefer fewer folds.
- ▶ If your model is a linear smoother there's a computational trick that makes LOO-CV extremely fast. (Problem Set)
- ▶ Asymptotic properties are related to  $K$  . . .

# Relationship between LOO-CV and TIC

## Theorem

LOO-CV using KL-divergence as the loss function is asymptotically equivalent to TIC but doesn't require us to estimate the Hessian and variance of the score.

# Large-sample Equivalence of LOO-CV and TIC

## Notation and Assumptions

For simplicity let  $Y_1, \dots, Y_T \sim \text{iid}$ . Let  $\hat{\theta}_{(t)}$  be the maximum likelihood estimator based on all observations **except**  $t$  and  $\hat{\theta}$  be the full-sample estimator.

## Log-likelihood as “Loss”

$CV_1 = \frac{1}{T} \sum_{t=1}^T \log f(y_t | \hat{\theta}_{(t)})$  but since min. KL = max. log-like.  
we choose the model with **highest**  $CV_1(m)$ .

# Overview of the Proof

First-Order Taylor Expansion of  $\log f(y_t|\hat{\theta}_{(t)})$  around  $\hat{\theta}$ :

$$\begin{aligned} CV_1 &= \frac{1}{T} \sum_{t=1}^T \log f(y_t|\hat{\theta}_{(t)}) \\ &= \frac{1}{T} \sum_{t=1}^T \left[ \log f(y_t|\hat{\theta}) + \frac{\partial \log f(y_t|\hat{\theta})}{\partial \theta'} (\hat{\theta}_{(t)} - \hat{\theta}) \right] + o_p(1) \\ &= \frac{\ell(\hat{\theta})}{T} + \frac{1}{T} \sum_{t=1}^T \frac{\partial \log f(y_t|\hat{\theta})}{\partial \theta'} (\hat{\theta}_{(t)} - \hat{\theta}) + o_p(1) \end{aligned}$$

Why isn't the first-order term zero in this case?



# Important Side Point

## Definition of ML Estimator

$$\frac{\partial \ell(\hat{\theta})}{\partial \theta'} = \frac{1}{T} \sum_{t=1}^T \frac{\partial \log f(y_t | \hat{\theta})}{\partial \theta} = 0$$

## In Contrast

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T \frac{\partial \log f(y_t | \hat{\theta})}{\partial \theta'} (\hat{\theta}_{(t)} - \hat{\theta}) &= \left[ \frac{1}{T} \sum_{t=1}^T \frac{\partial \log f(y_t | \hat{\theta})}{\partial \theta'} \hat{\theta}_{(t)} \right] - \hat{\theta} \left[ \frac{1}{T} \sum_{t=1}^T \frac{\partial \log f(y_t | \hat{\theta})}{\partial \theta'} \right] \\ &= \frac{1}{T} \sum_{t=1}^T \frac{\partial \log f(y_t | \hat{\theta})}{\partial \theta'} \hat{\theta}_{(t)} \neq 0 \end{aligned}$$

# Overview of Proof

From expansion two slides back, we simply need to show that:

$$\frac{1}{T} \sum_{t=1}^T \frac{\partial \log f(y_t | \hat{\theta})}{\partial \theta'} (\hat{\theta}_{(t)} - \hat{\theta}) = -\frac{1}{T} \text{tr}(\hat{J}^{-1} \hat{K}) + o_p(1)$$

$$\hat{K} = \frac{1}{T} \sum_{t=1}^T \left( \frac{\partial \log f(y_t | \hat{\theta})}{\partial \theta} \right) \left( \frac{\partial \log f(y_t | \hat{\theta})}{\partial \theta} \right)'$$

$$\hat{J} = -\frac{1}{T} \sum_{t=1}^T \frac{\partial^2 \log f(y_t | \hat{\theta})}{\partial \theta \partial \theta'}$$

## Overview of Proof

By the definition of  $\hat{K}$  and the properties of the trace operator:

$$\begin{aligned} -\frac{1}{T} \text{tr} \left\{ \hat{J}^{-1} \hat{K} \right\} &= -\frac{1}{T} \text{tr} \left\{ \hat{J}^{-1} \left[ \frac{1}{T} \sum_{t=1}^T \left( \frac{\partial \log f(y_t | \hat{\theta})}{\partial \theta} \right) \left( \frac{\partial \log f(y_t | \hat{\theta})}{\partial \theta} \right)' \right] \right\} \\ &= \left[ \frac{1}{T} \sum_{t=1}^T \text{tr} \left\{ -\frac{\hat{J}^{-1}}{T} \left( \frac{\partial \log f(y_t | \hat{\theta})}{\partial \theta} \right) \left( \frac{\partial \log f(y_t | \hat{\theta})}{\partial \theta} \right)' \right\} \right] \\ &= \frac{1}{T} \sum_{t=1}^T \frac{\partial \log f(y_t | \hat{\theta})}{\partial \theta'} \left( -\frac{1}{T} \hat{J}^{-1} \right) \frac{\partial \log f(y_t | \hat{\theta})}{\partial \theta} \end{aligned}$$

So it suffices to show that

$$\left( \hat{\theta}_{(t)} - \hat{\theta} \right) = -\frac{1}{T} \hat{J}^{-1} \left[ \frac{\partial \log f(y_t | \hat{\theta})}{\partial \theta} \right] + o_p(1)$$

# What is an Influence Function?

## Statistical Functional

$\mathbb{T} = \mathbb{T}(G)$  maps a CDF  $G$  to  $\mathbb{R}^p$ .

## Example: ML Estimation

$$\theta_0 = \mathbb{T}(G) = \arg \min_{\theta \in \Theta} E_G \left[ \log \left\{ \frac{g(Y)}{f(Y|\theta)} \right\} \right]$$

## Influence Function

Let  $\delta_y$  be a **point mass** at  $y$ :  $\delta_y(y) = 1$ ,  $\delta_y(y') = 0$  for  $y' \neq y$ .

Influence function = functional derivative: how does a small change in  $G$  affect  $\mathbb{T}$ ?

$$\text{infl}(G, y) = \lim_{\epsilon \rightarrow 0} \frac{\mathbb{T}[(1 - \epsilon) G + \epsilon \delta_y] - \mathbb{T}(G)}{\epsilon}$$

## Relating Influence Functions to $\hat{\theta}_{(t)}$

Empirical CDF  $\hat{G}$

$$\hat{G}(a) = \frac{1}{T} \sum_{t=1}^T \mathbf{1}\{y_t \leq a\} = \frac{1}{T} \sum_{t=1}^T \delta_{y_t}(a)$$

Relation to “LOO” Empirical CDF  $\hat{G}_{(t)}$

$$\hat{G} = \left(1 - \frac{1}{T}\right) \hat{G}_{(t)} + \frac{\delta_{y_t}}{T}$$

Applying  $\mathbb{T}$  to both sides...

$$\mathbb{T}(\hat{G}) = \mathbb{T}\left((1 - 1/T)\hat{G}_{(t)} + \delta_{y_t}/T\right)$$

## Relating Influence Functions to $\hat{\theta}_{(t)}$

Some algebra, followed by taking  $\varepsilon = 1/T$  to zero gives:

$$\mathbb{T}(\hat{G}) = \mathbb{T}\left((1 - 1/T)\hat{G}_{(t)} + \delta_{y_t}/T\right)$$

$$\mathbb{T}(\hat{G}) - \mathbb{T}(\hat{G}_{(t)}) = \mathbb{T}\left((1 - 1/T)\hat{G}_{(t)} + \delta_{y_t}/T\right) - \mathbb{T}(\hat{G}_{(t)})$$

$$\mathbb{T}(\hat{G}) - \mathbb{T}(\hat{G}_{(t)}) = \frac{1}{T} \left[ \frac{\mathbb{T}\left((1 - 1/T)\hat{G}_{(t)} + \delta_{y_t}/T\right) - \mathbb{T}(\hat{G}_{(t)})}{1/T} \right]$$

$$\mathbb{T}(\hat{G}) - \mathbb{T}(\hat{G}_{(t)}) = \frac{1}{T} \text{infl}\left(\hat{G}_{(t)}, y_t\right) + o_p(1)$$

$$\hat{\theta} - \hat{\theta}_{(t)} = \frac{1}{T} \text{infl}\left(\hat{G}, y_t\right) + o_p(1)$$

Last step: difference between having  $\hat{G}$  vs.  $\hat{G}_{(t)}$  in infl is negligible

# Steps for Last part of TIC/LOO-CV Equivalence Proof

## Step 1

Let  $\hat{G}$  denote the empirical CDF based on  $y_1, \dots, y_T$ . Then:

$$\left(\hat{\theta}_{(t)} - \hat{\theta}\right) = -\frac{1}{T} \text{infl}(\hat{G}, y_t) + o_p(1)$$

## Step 2

For ML estimation it turns out that:

$$\text{infl}(G, y) = J^{-1} \frac{\partial}{\partial \theta} \log f(y|\theta_0).$$

## Step 3

Evaluating Step 2 at  $\hat{G}$  and substituting into Step 2

$$\left(\hat{\theta}_{(t)} - \hat{\theta}\right) = -\frac{1}{T} \hat{J}^{-1} \left[ \frac{\partial \log f(y_t|\hat{\theta})}{\partial \theta} \right] + o_p(1)$$

# Derivation of Influence Function for MLE

I'll do this on the blackboard if we have time. . .



# Lecture #4 – Asymptotic Properties

Overview

Weak Consistency

Consistency

Efficiency

AIC versus BIC in a Simple Example

# Overview

## Asymptotic Properties

What happens as the sample size increases?

## Consistency

Choose “best” model with probability approaching 1 in the limit.

## Efficiency

Post-model selection estimator with low risk.

## Some References

Sin and White (1992, 1996), Pötscher (1991), Leeb & Pötscher (2005), Yang (2005) and Yang (2007).

# Penalizing the Likelihood

Examples we've seen:

$$TIC = 2\ell_T(\hat{\theta}) - \text{trace} \left\{ \hat{J}^{-1} \hat{K} \right\}$$

$$AIC = 2\ell_T(\hat{\theta}) - 2 \text{ length}(\theta)$$

$$BIC = 2\ell_T(\hat{\theta}) - \log(T) \text{ length}(\theta)$$

Generic penalty  $c_{T,k}$

$$IC(M_k) = 2 \sum_{t=1}^T \log f_{k,t}(Y_t | \hat{\theta}_k) - c_{T,k}$$

How does choice of  $c_{T,k}$  affect behavior of the criterion?

## Weak Consistency: Suppose $M_{k_0}$ Uniquely Minimizes KL

### Assumption

$$\liminf_{T \rightarrow \infty} \left( \min_{k \neq k_0} \frac{1}{T} \sum_{t=1}^T \{KL(g; f_{k,t}) - KL(g; f_{k_0,t})\} \right) > 0$$

### Consequences

- ▶ Any criterion with  $c_{T,k} > 0$  and  $c_{T,k} = o_p(T)$  is weakly consistent: **selects  $M_{k_0}$  wpa 1 in the limit.**
- ▶ Weak consistency still holds if  $c_{T,k}$  is zero for one of the models, so long as it is strictly positive for all the others.

## Both AIC and BIC are Weakly Consistent

Both satisfy  $T^{-1}c_{T,k} \xrightarrow{P} 0$ .

BIC Penalty:  $c_{T,k} = \log(T) \times \text{length}(\theta_k)$

AIC Penalty:  $c_{T,k} = 2 \times \text{length}(\theta_k)$

# Consistency: No Unique KL-minimizer

## Example

If the truth is an AR(5) model then AR(6), AR(7), AR(8), etc. models **all have zero KL-divergence**.

## Principle of Parsimony

Among the KL-minimizers, choose the **simplest model**, i.e. the one with the fewest parameters.

## Notation

$\mathcal{J}$  = be the set of all models that attain minimum KL-divergence

$\mathcal{J}_0$  = subset with the minimum number of parameters.

# Sufficient Conditions for Consistency

Consistency: Select Model from  $\mathcal{J}_0$  wpa 1

$$\lim_{T \rightarrow \infty} \mathbb{P} \left\{ \min_{\ell \in \mathcal{J} \setminus \mathcal{J}_0} [IC(M_{j_0}) - IC(M_\ell)] > 0 \right\} = 1$$

## Sufficient Conditions

(i) For all  $k \neq \ell \in \mathcal{J}$

$$\sum_{t=1}^T [\log f_{k,t}(Y_t | \theta_k^*) - \log f_{\ell,t}(Y_t | \theta_\ell^*)] = O_p(1)$$

where  $\theta_k^*$  and  $\theta_\ell^*$  are the KL minimizing parameter values.

(ii) For all  $j_0 \in \mathcal{J}_0$  and  $\ell \in (\mathcal{J} \setminus \mathcal{J}_0)$

$$P(c_{T,\ell} - c_{T,j_0} \rightarrow \infty) = 1$$

## BIC is Consistent; AIC and TIC Are Not

- ▶ AIC and TIC *cannot* satisfy (ii) since  $(c_{T,\ell} - c_{T,j_0})$  *does not depend on sample size*.
- ▶ It turns out that AIC and TIC are *not* consistent.
- ▶ BIC is consistent:

$$c_{T,\ell} - c_{T,j_0} = \log(T) \{ \text{length}(\theta_\ell) - \text{length}(\theta_{j_0}) \}$$

- ▶ Term in braces is *positive* since  $\ell \in \mathcal{J} \setminus \mathcal{J}_0$ , i.e.  $\ell$  is not as parsimonious as  $j_0$
- ▶  $\log(T) \rightarrow \infty$ , so BIC always selects a model in  $\mathcal{J}_0$  in the limit.



# Efficiency: Risk Properties of Post-selection Estimator

## Setup

- ▶ Models  $M_0$  and  $M_1$ ; corresponding estimators  $\hat{\theta}_{0,T}$  and  $\hat{\theta}_{1,T}$
- ▶ Model Selection: If  $\hat{M} = 0$  choose  $M_0$ ; if  $\hat{M} = 1$  choose  $M_1$ .

## Post-selection Estimator

$$\hat{\theta}_{\hat{M},T} \equiv \mathbf{1}_{\{\hat{M}=0\}} \hat{\theta}_{0,T} + \mathbf{1}_{\{\hat{M}=1\}} \hat{\theta}_{1,T}$$

## Two Sources of Randomness

Variability in  $\hat{\theta}_{\hat{M},T}$  arises both from  $(\hat{\theta}_{0,T}, \hat{\theta}_{1,T})$  and from  $\hat{M}$ .

## Question

How does the risk of  $\hat{\theta}_{\hat{M},T}$  compare to that of other estimators?

# Efficiency: Risk Properties of Post-selection Estimator

## Pointwise-risk Adaptivity

$\hat{\theta}_{\hat{M},T}$  is **pointwise-risk adaptive** if for any fixed  $\theta \in \Theta$ ,

$$\frac{R(\theta, \hat{\theta}_{\hat{M},T})}{\min \left\{ R(\theta, \hat{\theta}_{0,T}), R(\theta, \hat{\theta}_{1,T}) \right\}} \rightarrow 1, \quad \text{as } T \rightarrow \infty$$

## Minimax-rate Adaptivity

$\hat{\theta}_{\hat{M},T}$  is **minimax-rate adaptive** if

$$\sup_T \left[ \frac{\sup_{\theta \in \Theta} R(\theta, \hat{\theta}_{\hat{M},T})}{\inf_{\tilde{\theta}_T} \sup_{\theta \in \Theta} R(\theta, \tilde{\theta}_T)} \right] < \infty$$

# The Strengths of AIC and BIC Cannot be Shared

## Theorem

No model post-model selection estimator can be both pointwise-risk adaptive and minimax-rate adaptive.

## AIC vs. BIC

- ▶ BIC is pointwise-risk adaptive but AIC is not. (This is effectively identical to consistency.)
- ▶ AIC is minimax-rate adaptive, but BIC is not.
- ▶ Further Reading: Yang (2005), Yang (2007)

# Consistency and Efficiency in a Simple Example

## Information Criteria

Consider criteria of the form  $IC_m = 2\ell(\theta) - d_T \times \text{length}(\theta)$ .

## True DGP

$Y_1, \dots, Y_T \sim \text{iid } N(\mu, 1)$

## Candidate Models

$M_0$  assumes  $\mu = 0$ ,  $M_1$  does not restrict  $\mu$ . Only one parameter:

$$IC_0 = 2 \max_{\mu} \{\ell(\mu) : M_0\}$$

$$IC_1 = 2 \max_{\mu} \{\ell(\mu) : M_1\} - d_T$$

# Log-Likelihood Function

## Simple Algebra

$$\ell_T(\mu) = \text{Constant} - \frac{1}{2} \sum_{t=1}^T (Y_t - \mu)^2$$

## Tedious Algebra

$$\sum_{t=1}^T (Y_t - \mu)^2 = T(\bar{Y} - \mu)^2 + T\hat{\sigma}^2$$

## Combining These

$$\ell_T(\mu) = \text{Constant} - \frac{T}{2} (\bar{Y} - \mu)^2$$

# The Selected Model $\hat{M}$

## Information Criteria

$M_0$  sets  $\mu = 0$  while  $M_1$  uses the MLE  $\bar{Y}$ , so we have

$$IC_0 = 2 \max_{\mu} \{\ell(\mu) : M_0\} = 2 \times \text{Constant} - T\bar{Y}^2$$

$$IC_1 = 2 \max_{\mu} \{\ell(\mu) : M_1\} - d_T = 2 \times \text{Constant} - d_T$$

## Difference of Criteria

$$IC_1 - IC_0 = T\bar{Y}^2 - d_T$$

## Selected Model

$$\hat{M} = \begin{cases} M_1, & |\sqrt{T}\bar{Y}| \geq \sqrt{d_T} \\ M_0, & |\sqrt{T}\bar{Y}| < \sqrt{d_T} \end{cases}$$

## Verifying Weak Consistency: $\mu \neq 0$

KL Divergence for  $M_0$  and  $M_1$

$$KL(g; M_0) = \mu^2/2, \quad KL(g; M_1) = 0$$

Condition on KL-Divergence

$$\liminf_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \{KL(g; M_0) - KL(g; M_1)\} = \liminf_{n \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \left( \frac{\mu^2}{2} - 0 \right) > 0$$

Condition on Penalty

- ▶ Need  $c_{T,k} = o_p(T)$ , i.e.  $c_{T,k}/T \xrightarrow{P} 0$ .
- ▶ Both AIC and BIC satisfy this
- ▶ If  $\mu \neq 0$ , both AIC and BIC select  $M_1$  wpa 1 as  $T \rightarrow \infty$ .

## Verifying Consistency: $\mu = 0$

What's different?

- ▶ Both  $M_1$  and  $M_0$  are true and minimize KL divergence at zero.
- ▶ **Consistency** says choose most parsimonious true model:  $M_0$

## Verifying Conditions for Consistency

- ▶  $N(0, 1)$  model nested inside  $N(\mu, 1)$  model
- ▶ Truth is  $N(0, 1)$  so LR-stat is asymptotically  $\chi^2(1) = O_p(1)$ .
- ▶ For penalty term, need  $\mathbb{P}(c_{T,k} - c_{T,0}) \rightarrow \infty$
- ▶ BIC satisfies this but AIC doesn't.



# Finite-Sample Selection Probabilities: AIC

AIC Sets  $d_T = 2$

$$\hat{M}_{AIC} = \begin{cases} M_1, & |\sqrt{T}\bar{Y}| \geq \sqrt{2} \\ M_0, & |\sqrt{T}\bar{Y}| < \sqrt{2} \end{cases}$$

$$\begin{aligned} P(\hat{M}_{AIC} = M_1) &= P(|\sqrt{T}\bar{Y}| \geq \sqrt{2}) \\ &= P(|\sqrt{T}\mu + Z| \geq \sqrt{2}) \\ &= P(\sqrt{T}\mu + Z \leq -\sqrt{2}) + [1 - P(\sqrt{T}\mu + Z \leq \sqrt{2})] \\ &= \Phi(-\sqrt{2} - \sqrt{T}\mu) + [1 - \Phi(\sqrt{2} - \sqrt{T}\mu)] \end{aligned}$$

where  $Z \sim N(0, 1)$  since  $\bar{Y} \sim N(\mu, 1/T)$  because  $\text{Var}(Y_t) = 1$ .

# Finite-Sample Selection Probabilities: BIC

BIC sets  $d_T = \log(T)$

$$\hat{M}_{BIC} = \begin{cases} M_1, & |\sqrt{T}\bar{Y}| \geq \sqrt{\log(T)} \\ M_0, & |\sqrt{T}\bar{Y}| < \sqrt{\log(T)} \end{cases}$$

Same steps as for the AIC except with  $\sqrt{\log(T)}$  in the place of  $\sqrt{2}$ :

$$\begin{aligned} P(\hat{M}_{BIC} = M_1) &= P(|\sqrt{T}\bar{Y}| \geq \sqrt{\log(T)}) \\ &= \Phi(-\sqrt{\log(T)} - \sqrt{T}\mu) + [1 - \Phi(\sqrt{\log(T)} - \sqrt{T}\mu)] \end{aligned}$$

Interactive Demo: AIC vs BIC

[https://fditraglia.shinyapps.io/CH\\_Figure\\_4\\_1/](https://fditraglia.shinyapps.io/CH_Figure_4_1/)

# Probability of Over-fitting

- ▶ If  $\mu = 0$  both models are true but  $M_0$  is more parsimonious.
- ▶ Probability of over-fitting ( $Z$  denotes standard normal):

$$\begin{aligned}P(\hat{M} = M_1) &= P(|\sqrt{T}\bar{Y}| \geq \sqrt{d_T}) = P(|Z| \geq \sqrt{d_T}) \\&= P(Z^2 \geq d_T) = P(\chi_1^2 \geq d_T)\end{aligned}$$

- ▶ AIC:  $d_T = 2$  and  $P(\chi_1^2 \geq 2) \approx 0.157$ .
- ▶ BIC:  $d_T = \log(T)$  and  $P(\chi_1^2 \geq \log T) \rightarrow 0$  as  $T \rightarrow \infty$ .

AIC has  $\approx 16\%$  prob. of over-fitting; BIC does not over-fit in the limit.

# Risk of the Post-Selection Estimator

## The Post-Selection Estimator

$$\hat{\mu} = \begin{cases} \bar{Y}, & |\sqrt{T}\bar{Y}| \geq \sqrt{d_T} \\ 0, & |\sqrt{T}\bar{Y}| < \sqrt{d_T} \end{cases}$$

## Recall from above

Recall from above that  $\sqrt{T}\bar{Y} = \sqrt{T}\mu + Z$  where  $Z \sim N(0, 1)$

## Risk Function

MSE risk times  $T$  to get risk relative to minimax rate:  $1/T$ .

$$R(\mu, \hat{\mu}) = T \cdot \mathbb{E} \left[ (\hat{\mu} - \mu)^2 \right] = \mathbb{E} \left[ \left( \sqrt{T}\hat{\mu} - \sqrt{T}\mu \right)^2 \right]$$

# The Simplified MSE Risk Function

$$\begin{aligned}R(\mu, \hat{\mu}) &= 1 - [a\phi(a) - b\phi(b) + \Phi(b) - \Phi(a)] + T\mu^2 [\Phi(b) - \Phi(a)] \\ &= 1 + [b\phi(b) - a\phi(a)] + (T\mu^2 - 1) [\Phi(b) - \Phi(a)]\end{aligned}$$

where

$$\begin{aligned}a &= -\sqrt{d_T} - \sqrt{T}\mu \\ b &= \sqrt{d_T} - \sqrt{T}\mu\end{aligned}$$

[https://fditraglia.shinyapps.io/CH\\_Figure\\_4\\_2/](https://fditraglia.shinyapps.io/CH_Figure_4_2/)

# Understanding the Risk Plot

## AIC

- ▶ For any  $\mu \neq 0$ , risk  $\rightarrow 1$  as  $T \rightarrow \infty$ , the risk of the MLE
- ▶ For  $\mu = 0$ , risk  $\rightarrow 0$ , risk of “zero” estimator
- ▶ Max risk is bounded

## BIC

- ▶ For any  $\mu \neq 0$ , risk  $\rightarrow 1$  as  $T \rightarrow \infty$ , the risk of the MLE
- ▶ For  $\mu = 0$ , risk  $\rightarrow 0$ , risk of “zero” estimator
- ▶ Max risk is unbounded