

# 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  $\hat{\theta}$  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  $\alpha = \beta = \sqrt{n}/2$ ,

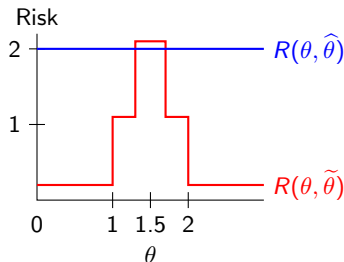
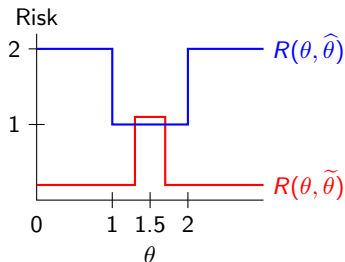
$$\hat{p} = \frac{n\bar{X} + \sqrt{n}/2}{n + \sqrt{n}}$$

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

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

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

$\text{KL}(g; f) \geq 0$  with equality iff  $g = f$  almost surely

### Jensen's Inequality

If  $\varphi$  is convex, then  $\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]$ , with strict equality when  $\varphi$  is affine or  $X$  is constant.

$\log$  is concave so  $(-\log)$  is convex

$$\begin{aligned}\mathbb{E}_G \left[ \log \left\{ \frac{g(Y)}{f(Y)} \right\} \right] &= \mathbb{E}_G \left[ -\log \left\{ \frac{f(Y)}{g(Y)} \right\} \right] \geq -\log \left\{ \mathbb{E}_G \left[ \frac{f(Y)}{g(Y)} \right] \right\} \\ &= -\log \left\{ \int_{-\infty}^{\infty} \frac{f(y)}{g(y)} \cdot g(y) dy \right\} \\ &= -\log \left\{ \int_{-\infty}^{\infty} f(y) dy \right\} \\ &= -\log(1) = 0\end{aligned}$$

# 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:  $RSS(m) = (\mathbf{y} - \mathbf{X}\hat{\beta}_m)'(\mathbf{y} - \mathbf{X}\hat{\beta}_m)$

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

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' Theorem 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...

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

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

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

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

Take Logs and Multiply by 2

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

The BIC

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

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

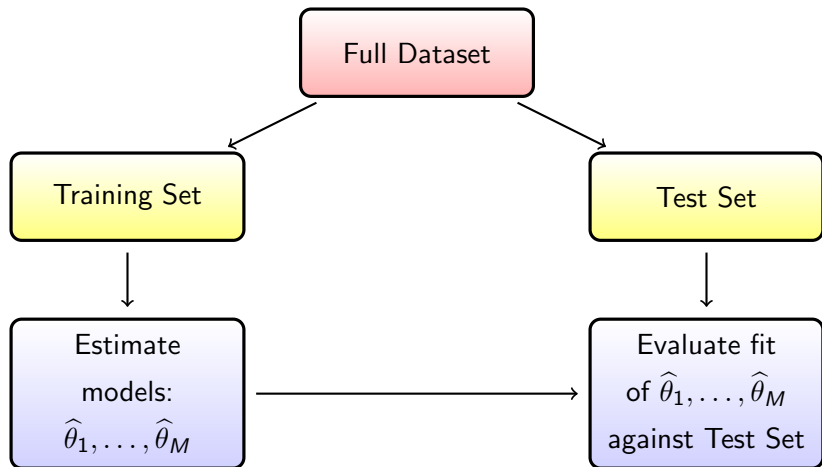
large-sample Frequentist approx. to Bayesian marginal likelihood



# Model Selection using a Hold-out Sample

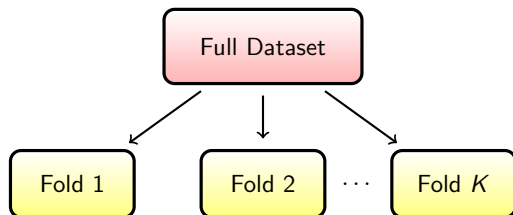
- ▶ The real problem is **double** use of the data: first for estimation, then for model comparison.
  - ▶ Maximized sample log-likelihood is an overly optimistic estimate of expected log-likelihood and hence KL-divergence
  - ▶ In-sample squared prediction error is an overly optimistic estimator of 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:

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

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

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

## 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

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

## Step 3

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

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



# 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

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

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

$$a = -\sqrt{d_T} - \sqrt{T}\mu$$

$$b = \sqrt{d_T} - \sqrt{T}\mu$$

[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

# Lecture #5 – Andrews (1999) Moment Selection Criteria

Lightning Review of GMM

The J-test Statistic Under Correct Specification

The J-test Statistic Under Mis-specification

Andrews (1999; Econometrica)

# Generalized Method of Moments (GMM) Estimation

## Notation

Let  $v_t$  be a  $(r \times 1)$  random vector,  $\theta$  be a  $(p \times 1)$  parameter vector, and  $f$  be a  $(q \times 1)$  vector of real-valued functions.

## Popn. Moment Conditions

$$\mathbb{E}[f(v_t, \theta_0)] = 0$$

## Sample Moment Conditions

$$\bar{g}_T(\theta) = \frac{1}{T} \sum_{t=1}^T f(v_t, \theta)$$

## GMM Estimator

$$\hat{\theta}_T = \arg \min_{\theta \in \Theta} \bar{g}_T(\theta)' \underset{(q \times q)}{W_T} \bar{g}_T(\theta), \quad W_T \rightarrow_p W \text{ (psd)}$$



# Key Assumptions for GMM I

## Stationarity

The sequence  $\{v_t: -\infty < t < \infty\}$  is strictly stationary. This implies that *any* functions of  $v_t$  are constant over  $t$ .

## Global Identification

$\mathbb{E}[f(v_t, \theta_0)] = 0$  but  $\mathbb{E}[f(v_t, \tilde{\theta})] \neq 0$  for any  $\tilde{\theta} \neq \theta_0$ .

## Regularity Conditions for Moment Functions

$f: \mathcal{V} \times \Theta \rightarrow \mathbb{R}^q$  satisfies:

- (i)  $f$  is  $v_t$ -almost surely continuous on  $\Theta$
- (ii)  $E[f(v_t, \theta)] < \infty$  exists and is continuous on  $\Theta$

# Key Assumptions for GMM I

## Regularity Conditions for Derivative Matrix

- (i)  $\nabla_{\theta'} f(v_t, \theta)$  exists and is  $v_t$ -almost continuous on  $\Theta$
- (ii)  $E[\nabla_{\theta} f(v_t, \theta_0)] < \infty$  exists and is continuous in a neighborhood  $N_{\epsilon}$  of  $\theta_0$
- (iii)  $\sup_{\theta \in N_{\epsilon}} \left\| T^{-1} \sum_{t=1}^T \nabla_{\theta} f(v_t, \theta) - E[\nabla_{\theta} f(v_t, \theta)] \right\| \xrightarrow{P} 0$

## Regularity Conditions for Variance of Moment Conditions

- (i)  $E[f(v_t, \theta_0)f(v_t, \theta_0)']$  exists and is finite.
- (ii)  $\lim_{T \rightarrow \infty} \text{Var} \left[ \sqrt{T} \bar{g}_T(\theta_0) \right] = S$  exists and is a finite, positive definite matrix.

# Main Results for GMM Estimation

Under the Assumptions Described Above

Consistency:  $\hat{\theta}_T \xrightarrow{p} \theta_0$

Asymptotic Normality:  $\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{d} \mathcal{N}(0, MSM')$

$$M = (G_0'WG_0)^{-1}G_0'W$$

$$S = \lim_{T \rightarrow \infty} \text{Var} \left[ \sqrt{T} \bar{g}_T(\theta_0) \right]$$

$$G_0 = E[\nabla_{\theta'} f(v_t, \theta_0)]$$

$$W = \text{plim}_{T \rightarrow \infty} W_T$$

# The J-test Statistic

$$J_T = T \bar{g}_T(\hat{\theta}'_T) \hat{S}^{-1} \bar{g}_T(\hat{\theta}_T)$$

$$\hat{S} \rightarrow_p S = \lim_{T \rightarrow \infty} \text{Var} \left[ \sqrt{T} \bar{g}_T(\theta_0) \right]$$

$$\bar{g}_T(\hat{\theta}_T) = \frac{1}{T} \sum_{t=1}^T f(v_t, \hat{\theta}_T)$$

$$\hat{\theta}_T = \text{GMM Estimator}$$

## Case I: Correct Specification

Suppose that all of the preceding assumptions hold, in particular that the model is **correctly specified**:

$$\mathbb{E}[f(v_t, \theta_0)] = 0$$

Recall that under the standard assumptions, the GMM estimator is consistent **regardless of the choice of  $W_T$** ...

## Case I: Taylor Expansion under Correct Specification

$$W_T^{1/2} \sqrt{T} \bar{g}_T(\hat{\theta}_T) = [I_q - P(\theta_0)] W_T^{1/2} \sqrt{T} \bar{g}_T(\theta_0) + o_p(1)$$

$$P(\theta_0) = F(\theta_0) [F(\theta_0)' F(\theta_0)]^{-1} F(\theta_0)'$$

$$F(\theta_0) = W_T^{1/2} E[\nabla_{\theta} f(v_t, \theta_0)]$$

### Over-identification

If  $\dim(f) > \dim(\theta_0)$ ,  $W_T^{1/2} \mathbb{E}[f(v_t, \theta_0)]$  is the linear combn. used in GMM estimation.

### Identifying and Over-Identifying Restrictions

$P(\theta_0) \equiv$  **identifying restrictions**;

$I_q - P(\theta_0) \equiv$  **over-identifying restrictions**

## J-test Statistic Under Correct Specification

$$W_T^{1/2} \sqrt{T} \bar{g}_T(\hat{\theta}_T) = [I_q - P(\theta_0)] W_T^{1/2} \sqrt{T} \bar{g}_T(\theta_0) + o_p(1)$$

- ▶ CLT for  $\sqrt{T} \bar{g}_T(\theta_0)$
- ▶  $I_q - P(\theta_0)$  has rank  $(q - p)$ , since  $P(\theta_0)$  has rank  $p$ .
- ▶ **Singular** normal distribution
- ▶  $W_T^{1/2} \sqrt{T} \bar{g}_T(\hat{\theta}_T) \xrightarrow{d} \mathcal{N}(0, N W_T^{1/2} S W_T^{1/2} N')$
- ▶ Substituting  $\hat{S}^{-1}$ ,  $J_T \xrightarrow{d} \chi_{q-p}^2$

## Case II: Fixed Mis-specification

$$\mathbb{E}[f(v_t, \theta)] = \mu(\theta), \quad \|\mu(\theta)\| > 0, \quad \forall \theta \in \Theta$$

N.B.

This can *only* occur in the over-identified case, since we can always solve the population moment conditions in the just-identified case.

Notation

- ▶  $\theta^* \equiv$  solution to identifying restrictions ( $\hat{\theta}_T \rightarrow_p \theta^*$ )
- ▶  $\mu^* = \mu(\theta^*) = \text{plim}_{T \rightarrow \infty} \bar{g}_T(\hat{\theta}_T)$



## Case II: Fixed Mis-specification

$$\frac{1}{T} J_T = \bar{g}_T(\hat{\theta}_T)' \hat{S}^{-1} \bar{g}_T(\hat{\theta}_T) = \mu_*' W \mu_* + o_p(1)$$

- ▶  $W$  positive definite
- ▶ since  $\mu(\theta) > 0$  for all  $\theta \in \Theta$ .
- ▶ Hence:  $\mu_*' W \mu_* > 0$
- ▶ Fixed mis-specification  $\Rightarrow J$ -test statistic *diverges at rate  $T$* :

$$J_T = T \mu_*' W \mu_* + o_p(T)$$

## Summary: Correct Specification vs. Fixed Mis-specification

Correct Specification:  $J_T \Rightarrow \chi^2_{q-p} = O_p(1)$

Fixed Mis-specification:  $J_T = O_p(T)$

## Andrews (1999; Econometrica)

- ▶ Family of moment selection criteria (MSC) for GMM
- ▶ Aims to **consistently** choose any and all correct MCs and eliminate incorrect MCs
- ▶ AIC/BIC: add a **penalty** to maximized log-likelihood
- ▶ Andrews MSC: add a **bonus** term to the J-statistic
  - ▶ J-stat shows how well MCs “fit”
  - ▶ Compares  $\hat{\theta}_T$  estimated using  $P(\theta_0)$  to MCs from  $I_q - P(\theta_0)$
  - ▶ J-stat tends to increase with degree of overidentification even if MCs are correct, since it converges to a  $\chi^2_{q-p}$

## Andrews (1999) – Notation

$f_{max} \equiv (q \times 1)$  vector of all MCs under consideration

$c \equiv (q \times 1)$  selection vector: zeros and ones indicating which MCs are included

$\mathcal{C} \equiv$  set of all candidates  $c$

$|c| \equiv \#$  of MCs in candidate  $c$

Let  $\hat{\theta}_T(c)$  be the efficient two-step GMM estimator based on the moment conditions  $E[f(v_t, \theta, c)] = 0$  and define

$$V_{\theta}(c) = \left[ G_0(c) S(c)^{-1} G_0(c) \right]^{-1}$$

$$G_0(c) = E[\nabla'_{\theta} f(v_t, \theta_0; c)]$$

$$S(c) = \lim_{T \rightarrow \infty} \text{Var} \left[ \frac{1}{\sqrt{T}} \sum_{t=1}^T f(v_t, \theta_0; c) \right]$$

$$J_T(c) = T \bar{g}_T \left( \hat{\theta}_T(c); c \right)' \hat{S}_T(c)^{-1} \bar{g}_T \left( \hat{\theta}_T(c); c \right)$$

# Identification Condition

- ▶ Andrews wants maximal set of correct MCs
  - ▶ Consistent, minimum asymptotic variance
- ▶ But different  $\theta$  values could solve  $\mathbb{E}[f(v_t, \theta, c)]$  for different  $c$ !
- ▶ Which  $\theta_0$  are we actually trying to be consistent for?

## More Notation

- ▶  $\mathcal{Z}^0 \equiv$  set of all  $c$  for which  $\exists \theta$  with  $\mathbb{E}[f(v_t, \theta, c)] = 0$
- ▶  $\mathcal{MZ}^0 \equiv$  subset of  $\mathcal{Z}^0$  with **maximal**  $|c|$ .

## Assumption

Andrews assumes that  $\mathcal{MZ}^0 = \{c_0\}$ , a singleton.

# Family of Moment Selection Criteria

- ▶ Criteria of the form  $MSC(c) = J_T(c) - B(T, |c|)$
- ▶  $B$  is a **bonus term** that depends on sample size and # of MCs
- ▶ Choose  $\hat{c}_T = \arg \min_{c \in \mathcal{C}} MSC(c)$
- ▶ Implementation Detail: Andrews suggests using a **centered** covariance matrix estimator:

$$\hat{S}(c) = \frac{1}{T} \sum_{t=1}^T \left[ f(v_t, \hat{\theta}_T(c); c) - \bar{g}_T(\hat{\theta}_T(c); c) \right] \left[ f(v_t, \hat{\theta}_T(c); c) - \bar{g}_T(\hat{\theta}_T(c); c) \right]'$$

based on the weighting matrix that *would be* efficient if the moment conditions were correctly specified. This remains consistent for  $S(c)$  even under fixed mis-specification

## Regularity Conditions for the $J$ -test Statistic

- (i) If  $\mathbb{E}[f(v_t, \theta; c)] = 0$  for a unique  $\theta \in \Theta$ , then  $J_T(c) \xrightarrow{d} \chi^2_{|c|-p}$
- (ii) If  $\mathbb{E}[f(v_t, \theta; c)] \neq 0$  for a *all*  $\theta \in \Theta$  then  $T^{-1}J_T(c) \xrightarrow{p} a(c)$ , a finite, positive constant that may depend on  $c$ .

## Regularity Conditions for Bonus Term

The bonus term can be written as  $B(|c|, T) = \kappa_T h(|c|)$ , where

- (i)  $h(\cdot)$  is strictly increasing
- (ii)  $\kappa_T \rightarrow \infty$  as  $T \rightarrow \infty$  and  $\kappa_T = o(T)$

## Identification Conditions

- (i)  $\mathcal{M}\mathcal{Z}^0 = \{c_0\}$
- (ii)  $\mathbb{E}[f(v_t, \theta_0; c_0)] = 0$  and  $E[f(v_t, \theta; c_0)] \neq 0$  for any  $\theta \neq \theta_0$

# Consistency of Moment Selection

## Theorem

Under the preceding assumptions,  $MSC(c)$  is a consistent moment selection criterion, i.e.  $\hat{c}_T \xrightarrow{P} c_0$ .

## Some Examples

$$\text{GMM-BIC}(c) = J_T(c) - (|c| - p) \log(T)$$

$$\text{GMM-HQ}(c) = J_T(c) - 2.01 (|c| - p) \log(\log(T))$$

$$\text{GMM-AIC}(c) = J_T(c) - 2 (|c| - p)$$

## How do these examples behave?

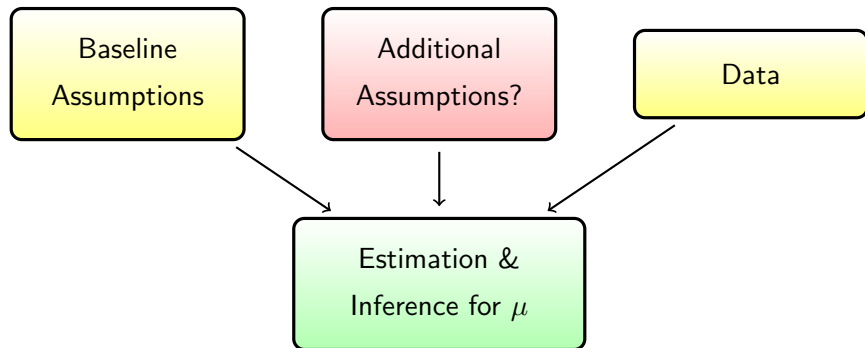
GMM-AIC doesn't satisfy the theorem, since  $\kappa_T = 2$  does not diverge as  $T \rightarrow \infty$ . GMM-BIC and GMM-HQ are consistent since  $\lim_{T \rightarrow \infty} \log(T)/T = 0$  and  $\lim_{T \rightarrow \infty} \log(\log(T))/T = 0$ .



# Lecture #6 – Focused Moment Selection

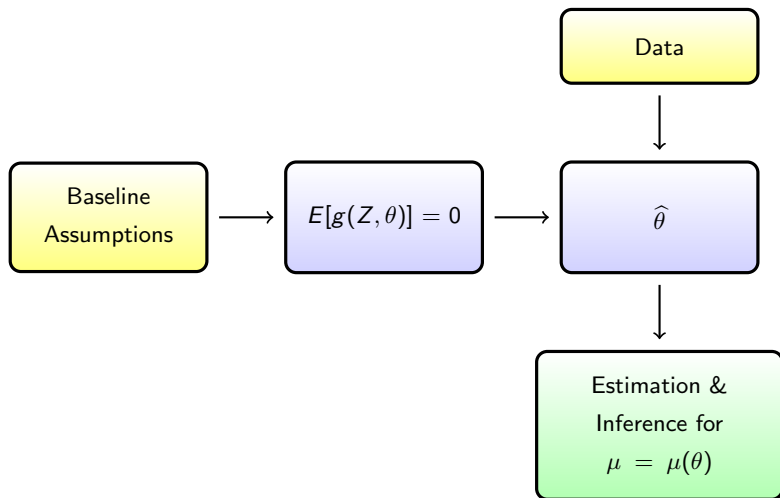
DiTraglia (2016, JoE)

# Focused Moment Selection Criterion (FMSC)

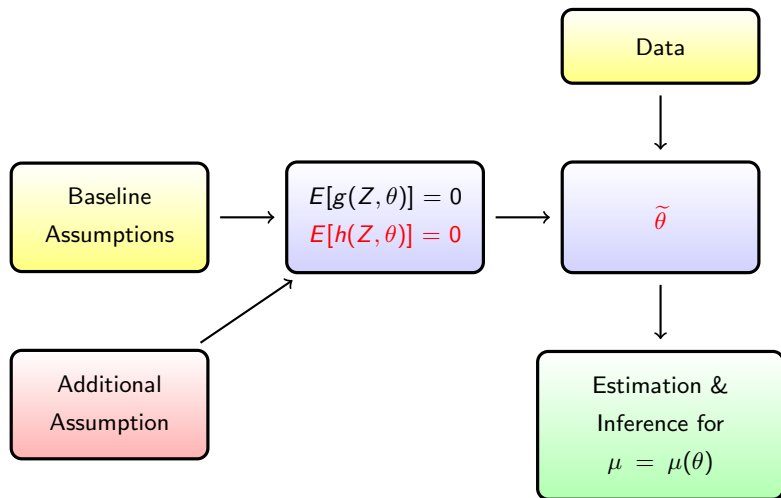


1. Choose False Assumptions on Purpose
2. Focused Choice of Assumptions
3. Local mis-specification
4. Averaging, Inference post-selection

# GMM Framework



# Adding Moment Conditions



# Ordinary versus Two-Stage Least Squares

$$y_i = \beta x_i + \epsilon_i$$

$$x_i = \mathbf{z}_i' \boldsymbol{\pi} + v_i$$

$$E[\mathbf{z}_i \epsilon_i] = 0$$

$$E[x_i \epsilon_i] = ?$$

## Choosing Instrumental Variables

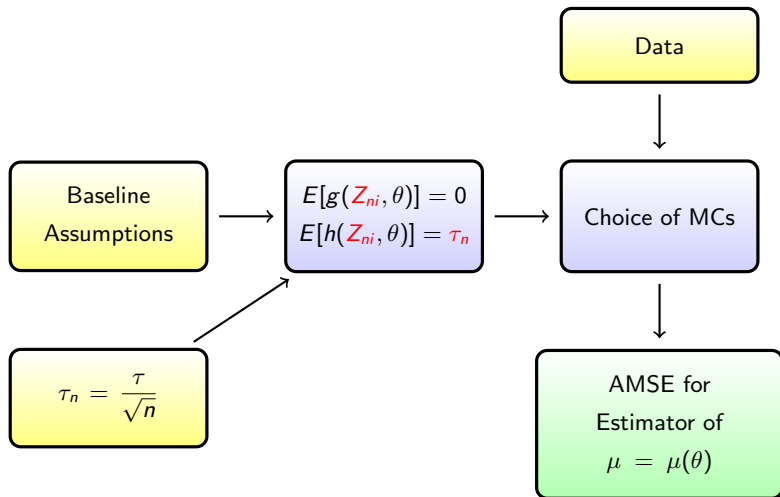
$$y_i = \beta x_i + \epsilon_i$$

$$x_i = \Pi'_1 \mathbf{z}_i^{(1)} + \Pi'_2 \mathbf{z}_i^{(2)} + v_i$$

$$E[\mathbf{z}_i^{(1)} \epsilon_i] = 0$$

$$E[\mathbf{z}_i^{(2)} \epsilon_i] = ?$$

# FMSC Asymptotics – Local Mis-Specification



# Local Mis-Specification for OLS versus TSLS

$$y_i = \beta x_i + \epsilon_i$$

$$x_i = \mathbf{z}_i' \boldsymbol{\pi} + v_i$$

$$E[\mathbf{z}_i \epsilon_i] = 0$$

$$E[x_i \epsilon_i] = \tau / \sqrt{n}$$



## Local Mis-Specification for Choosing IVs

$$\begin{aligned}y_i &= \beta x_i + \epsilon_i \\x_i &= \Pi'_1 \mathbf{z}_i^{(1)} + \Pi'_2 \mathbf{z}_i^{(2)} + v_i\end{aligned}$$

$$E[\mathbf{z}_i^{(1)} \epsilon_i] = 0$$

$$E[\mathbf{z}_i^{(2)} \epsilon_i] = \tau / \sqrt{n}$$

# Local Mis-Specification

Triangular Array  $\{Z_{ni}: 1 \leq i \leq n, n = 1, 2, \dots\}$  with

(a)  $E[g(Z_{ni}, \theta_0)] = 0$

(b)  $E[h(Z_{ni}, \theta_0)] = n^{-1/2}\tau$

(c)  $\{f(Z_{ni}, \theta_0): 1 \leq i \leq n, n = 1, 2, \dots\}$  uniformly integrable

(d)  $Z_{ni} \rightarrow_d Z_i$ , where the  $Z_i$  are identically distributed.

Shorthand: Write  $Z$  for  $Z_i$

## Candidate GMM Estimator

$$\hat{\theta}_S = \arg \min_{\theta \in \Theta} [\Xi_S f_n(\theta)]' \widetilde{W}_S [\Xi_S f_n(\theta)]$$

$\Xi_S$  = Selection Matrix (ones and zeros)

$\widetilde{W}_S$  = Weight Matrix (p.s.d.)

$$f_n(\theta) = \begin{bmatrix} g_n(\theta) \\ h_n(\theta) \end{bmatrix} = \begin{bmatrix} n^{-1} \sum_{i=1}^n g(Z_{ni}, \theta) \\ n^{-1} \sum_{i=1}^n h(Z_{ni}, \theta) \end{bmatrix}$$

## Notation: Limit Quantities

$$G = E [\nabla_{\theta} g(Z, \theta_0)], \quad H = E [\nabla_{\theta} h(Z, \theta_0)], \quad F = \begin{bmatrix} G \\ H \end{bmatrix}$$

$$\Omega = \text{Var} [f(Z, \theta_0)] = \begin{bmatrix} \Omega_{gg} & \Omega_{gh} \\ \Omega_{hg} & \Omega_{hh} \end{bmatrix}$$

$$\widetilde{W}_S \rightarrow_p W_S \text{ (p.d.)}$$

# Local Mis-Specification + Standard Regularity Conditions

Every candidate estimator is consistent for  $\theta_0$  and

$$\sqrt{n}(\hat{\theta}_S - \theta_0) \rightarrow_d -K_S \Xi_S \left( \begin{bmatrix} M_g \\ M_h \end{bmatrix} + \begin{bmatrix} 0 \\ \tau \end{bmatrix} \right)$$

$$K_S = [F_S' W_S F_S]^{-1} F_S' W_S$$

$$M = (M_g', M_h')'$$

$$M \sim N(0, \Omega)$$

## Scalar Target Parameter $\mu$

$$\mu = \mu(\theta) \quad \text{Z-a.s. continuous function}$$

$$\mu_0 = \mu(\theta_0) \quad \text{true value}$$

$$\hat{\mu}_S = \mu(\hat{\theta}_S) \quad \text{estimator}$$

## Delta Method

$$\sqrt{n}(\hat{\mu}_S - \mu_0) \rightarrow_d -\nabla_{\theta}\mu(\theta_0)'K_S\Xi_S \left( M + \begin{bmatrix} 0 \\ \tau \end{bmatrix} \right)$$

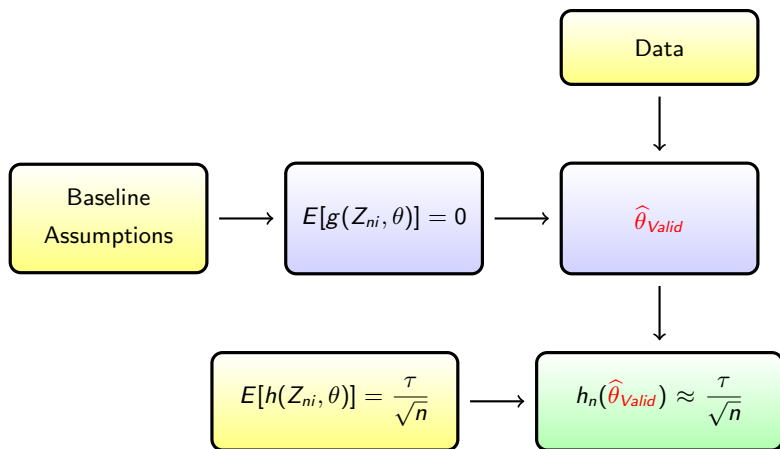
FMSC: Estimate  $AMSE(\hat{\mu}_S)$  and minimize over  $S$

$$AMSE(\hat{\mu}_S) = \nabla_{\theta}\mu(\theta_0)' K_S \Xi_S \left\{ \begin{bmatrix} 0 & 0 \\ 0 & \tau\tau' \end{bmatrix} + \Omega \right\} \Xi_S' K_S' \nabla_{\theta}\mu(\theta_0)$$

Estimating the unknowns

No consistent estimator of  $\tau$  exists! (But everything else is easy)

# A Plug-in Estimator of $\tau$





## An Asymptotically Unbiased Estimator of $\tau\tau'$

$$\sqrt{nh_n}(\hat{\theta}_v) = \hat{\tau} \rightarrow_d (\Psi M + \tau) \sim N_q(\tau, \Psi\Omega\Psi')$$

$$\Psi = \begin{bmatrix} -HK_v & \mathbf{I}_q \end{bmatrix}$$

$\hat{\tau}\hat{\tau}' - \hat{\Psi}\hat{\Omega}\hat{\Psi}$  is an asymptotically unbiased estimator of  $\tau\tau'$ .

## FMSC: Asymptotically Unbiased Estimator of AMSE

$$\text{FMSC}_n(S) = \nabla_{\theta} \mu(\hat{\theta})' \hat{K}_S \Xi_S \left\{ \begin{bmatrix} 0 & 0 \\ 0 & \hat{B} \end{bmatrix} + \hat{\Omega} \right\} \Xi_S' \hat{K}_S' \nabla_{\theta} \mu(\hat{\theta})$$

$$\hat{B} = \hat{\tau} \hat{\tau}' - \hat{\psi} \hat{\Omega} \hat{\psi}'$$

Choose  $S$  to minimize  $\text{FMSC}_n(S)$  over the set of candidates  $\mathcal{S}$ .

## A (Very) Special Case of the FMSC

Under homoskedasticity, FMSC selection in the OLS versus TSLS example is *identical* to a Durbin-Hausman-Wu test with  $\alpha \approx 0.16$

$$\hat{\tau} = n^{-1/2} \mathbf{x}'(\mathbf{y} - \mathbf{x}\tilde{\beta}_{TSLS})$$

OLS gets benefit of the doubt, but not as much as  $\alpha = 0.05, 0.1$

# Limit Distribution of FMSC

$FMSC_n(S) \rightarrow_d FMSC_S$ , where

$$\begin{aligned} FMSC_S &= \nabla_{\theta}\mu(\theta_0)' K_S \Xi_S \left\{ \begin{bmatrix} 0 & 0 \\ 0 & B \end{bmatrix} + \Omega \right\} \Xi_S' K_S' \nabla_{\theta}\mu(\theta_0) \\ B &= (\Psi M + \tau)(\Psi M + \tau)' - \Psi \Omega \Psi' \end{aligned}$$

*Conservative criterion: random even in the limit.*

# Moment Average Estimators

$$\hat{\mu} = \sum_{S \in \mathcal{S}} \hat{w}_S \hat{\mu}_S$$

## Additional Notation

$\hat{\mu}$  Moment-average Estimator

$\hat{\mu}_S$  Estimator of target parameter under moment set  $S$

$\hat{w}_S$  Data-dependent weight function

$\mathcal{S}$  Collection of moment sets under consideration

# Examples of Moment-Averaging Weights

## Post-Moment Selection Weights

$$\hat{\omega}_S = \mathbf{1} \{ \text{MSC}_n(S) = \min_{S' \in \mathcal{S}} \text{MSC}_n(S') \}$$

## Exponential Weights

$$\hat{\omega}_S = \exp \left\{ -\frac{\kappa}{2} \text{MSC}(S) \right\} / \sum_{S' \in \mathcal{S}} \exp \left\{ -\frac{\kappa}{2} \text{MSC}(S') \right\}$$

## Minimum-AMSE Weights...

## Minimum AMSE-Averaging Estimator: OLS vs. TSLS

$$\tilde{\beta}(\omega) = \omega \hat{\beta}_{OLS} + (1 - \omega) \tilde{\beta}_{TSLS}$$

Under homoskedasticity:

$$\omega^* = \left[ 1 + \frac{\text{ABIAS(OLS)}^2}{\text{AVAR(TSLS)} - \text{AVAR(OLS)}} \right]^{-1}$$

Estimate by:

$$\hat{\omega}^* = \left[ 1 + \frac{\max \{0, (\hat{\tau}^2 - \hat{\sigma}_\epsilon^2 \hat{\sigma}_x^2 (\hat{\sigma}_x^2 / \hat{\gamma}^2 - 1)) / \hat{\sigma}_x^4\}}{\hat{\sigma}_\epsilon^2 (1 / \hat{\gamma}^2 - 1 / \hat{\sigma}_x^2)} \right]^{-1}$$

Where  $\hat{\gamma}^2 = n^{-1} \mathbf{x}' Z (Z' Z)^{-1} Z' \mathbf{x}$

# Limit Distribution of Moment-Average Estimators

$$\hat{\mu} = \sum_{S \in \mathcal{S}} \hat{\omega}_S \hat{\mu}_S$$

- (i)  $\sum_{S \in \mathcal{S}} \hat{\omega}_S = 1$  a.s.
- (ii)  $\hat{\omega}(S) \rightarrow_d \varphi_S(\tau, M)$  a.s.-continuous function of  $\tau$ ,  $M$  and consistently-estimable constants only

$$\sqrt{n}(\hat{\mu} - \mu_0) \rightarrow_d \Lambda(\tau)$$

$$\Lambda(\tau) = -\nabla_{\theta} \mu(\theta_0)' \left[ \sum_{S \in \mathcal{S}} \varphi_S(\tau, M) K_S \Xi_S \right] \left( M + \begin{bmatrix} 0 \\ \tau \end{bmatrix} \right)$$



# Simulating from the Limit Experiment

Suppose  $\tau$  Known, Consistent Estimators of Everything Else

- for  $j \in \{1, 2, \dots, J\}$ 
  - $M_j \stackrel{iid}{\sim} N_{p+q} \left( 0, \hat{\Omega} \right)$
  - $\Lambda_j(\tau) = -\nabla_{\theta} \mu(\hat{\theta})' \left[ \sum_{s \in \mathcal{S}} \hat{\varphi}_s(M_j + \tau) \hat{K}_s \Xi_s \right] (M_j + \tau)$
- Using  $\{\Lambda_j(\tau)\}_{j=1}^J$  calculate  $\hat{a}(\tau)$ ,  $\hat{b}(\tau)$  such that
$$P \left[ \hat{a}(\tau) \leq \Lambda(\tau) \leq \hat{b}(\tau) \right] = 1 - \alpha$$
- $P \left[ \hat{\mu} - \hat{b}(\tau)/\sqrt{n} \leq \mu_0 \leq \hat{\mu} - \hat{a}(\tau)/\sqrt{n} \right] \approx 1 - \alpha$

## Two-step Procedure for Conservative Intervals

1. Construct  $1 - \delta$  confidence region  $\mathcal{T}(\hat{\tau}, \delta)$  for  $\tau$
2. For each  $\tau^* \in \mathcal{T}(\hat{\tau}, \delta)$  calculate  $1 - \alpha$  confidence interval  $[\hat{a}(\tau^*), \hat{b}(\tau^*)]$  for  $\Lambda(\tau^*)$  as described on previous slide.
3. Take the lower and upper bound over the resulting intervals:  
 $\hat{a}_{min}(\hat{\tau}) = \min_{\tau^* \in \mathcal{T}} \hat{a}(\tau^*), \quad \hat{b}_{max}(\hat{\tau}) = \max_{\tau^* \in \mathcal{T}} \hat{b}(\tau^*)$
4. The interval

$$CI_{sim} = \left[ \hat{\mu} - \frac{\hat{b}_{max}(\hat{\tau})}{\sqrt{n}}, \quad \hat{\mu} - \frac{\hat{a}_{min}(\hat{\tau})}{\sqrt{n}} \right]$$

has asymptotic coverage of at least  $1 - (\alpha + \delta)$

## OLS versus TSLS Simulation

$$y_i = 0.5x_i + \epsilon_i$$

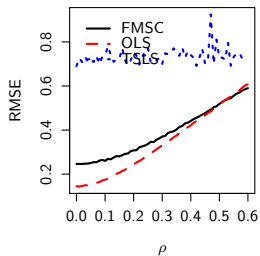
$$x_i = \pi(z_{1i} + z_{2i} + z_{3i}) + v_i$$

$$(\epsilon_i, v_i, z_{1i}, z_{2i}, z_{3i}) \sim \text{iid } N(0, \mathcal{S})$$

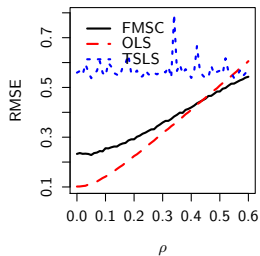
$$\mathcal{S} = \begin{bmatrix} 1 & \rho & 0 & 0 & 0 \\ \rho & 1 - \pi^2 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/3 \end{bmatrix}$$

$$\text{Var}(x) = 1, \quad \rho = \text{Cor}(x, \epsilon), \quad \pi^2 = \text{First-Stage } R^2$$

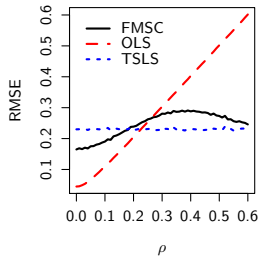
$N = 50, \pi = 0.2$



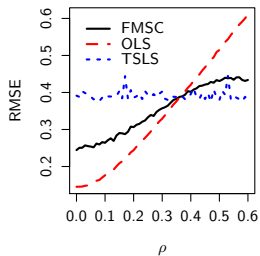
$N = 100, \pi = 0.2$



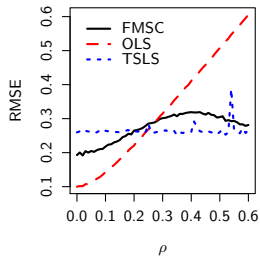
$N = 500, \pi = 0.2$



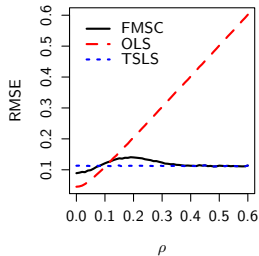
$N = 50, \pi = 0.4$



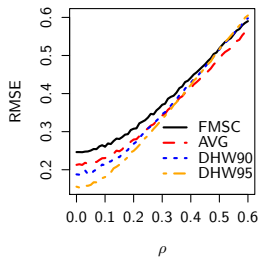
$N = 100, \pi = 0.4$



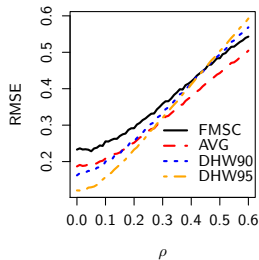
$N = 500, \pi = 0.4$



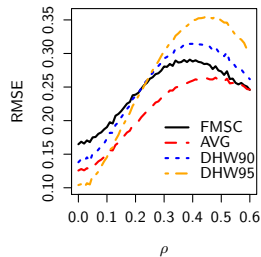
$N = 50, \pi = 0.2$



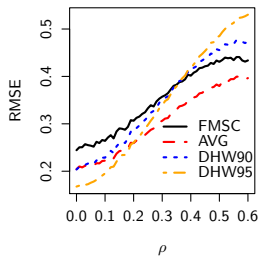
$N = 100, \pi = 0.2$



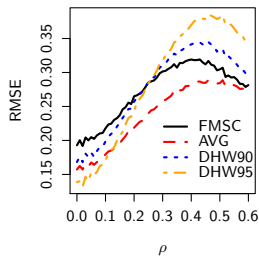
$N = 500, \pi = 0.2$



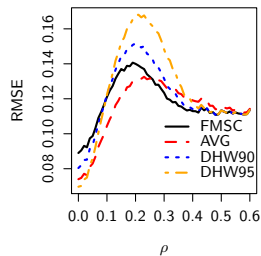
$N = 50, \pi = 0.4$



$N = 100, \pi = 0.4$



$N = 500, \pi = 0.4$



## Choosing Instrumental Variables Simulation

$$y_i = 0.5x_i + \epsilon_i$$

$$x_i = (z_{1i} + z_{2i} + z_{3i})/3 + \gamma w_i + v_i$$

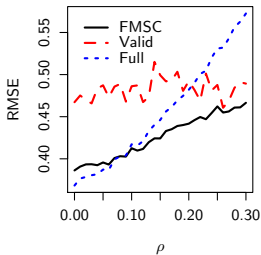
$$(\epsilon_i, v_i, w_i, z_{1i}, z_{2i}, z_{3i})' \sim \text{iid } N(0, \mathcal{V})$$

$$\mathcal{V} = \begin{bmatrix} 1 & (0.5 - \gamma\rho) & \rho & 0 & 0 & 0 \\ (0.5 - \gamma\rho) & (8/9 - \gamma^2) & 0 & 0 & 0 & 0 \\ \rho & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/3 \end{bmatrix}$$

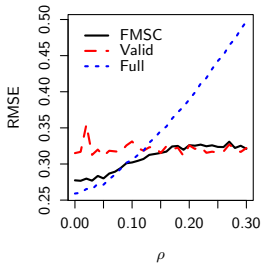
$$\gamma = \text{Cor}(x, w), \quad \rho = \text{Cor}(w, \epsilon), \quad \text{First-Stage } R^2 = 1/9 + \gamma^2$$

$$\text{Var}(x) = 1, \quad \text{Cor}(x, \epsilon) = 0.5$$

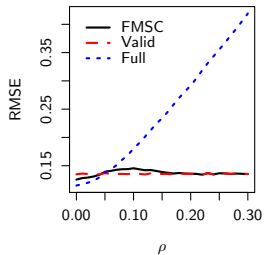
$N = 50, \gamma = 0.2$



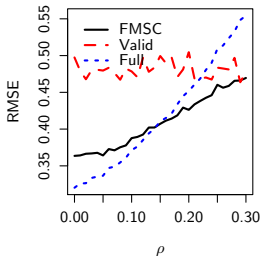
$N = 100, \gamma = 0.2$



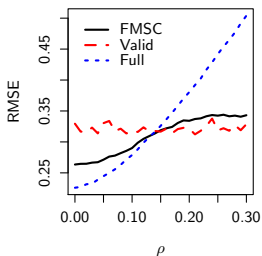
$N = 500, \gamma = 0.2$



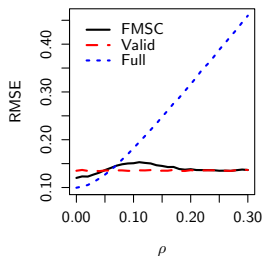
$N = 50, \gamma = 0.3$



$N = 100, \gamma = 0.3$



$N = 500, \gamma = 0.3$



# Alternative Moment Selection Procedures

## Downward $J$ -test

Use Full instrument set unless  $J$ -test rejects.

## Andrews (1999) – GMM Moment Selection Criteria

$$\text{GMM-MS}(S) = J_n(S) - \text{Bonus}$$

## Hall & Peixe (2003) – Canonical Correlations Info. Criterion

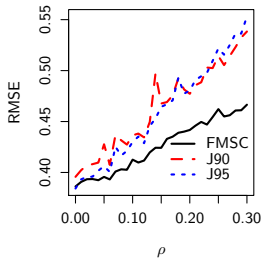
$$\text{CCIC}(S) = n \log [1 - R_n^2(S)] + \text{Penalty}$$

## Penalty/Bonus Terms

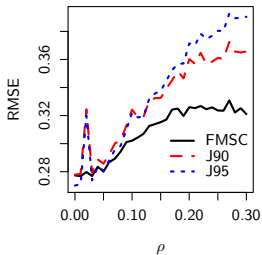
Analogies to AIC, BIC, and Hannan-Quinn



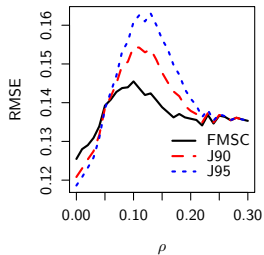
$N = 50, \gamma = 0.2$



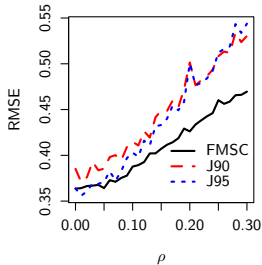
$N = 100, \gamma = 0.2$



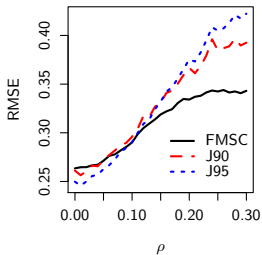
$N = 500, \gamma = 0.2$



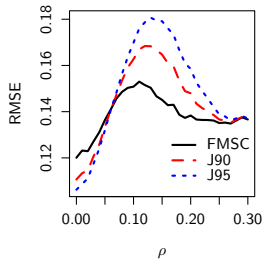
$N = 50, \gamma = 0.3$



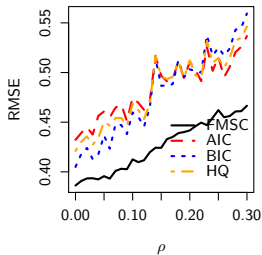
$N = 100, \gamma = 0.3$



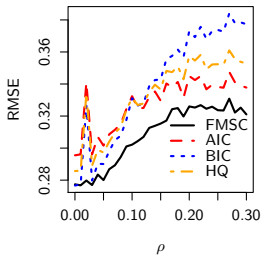
$N = 500, \gamma = 0.3$



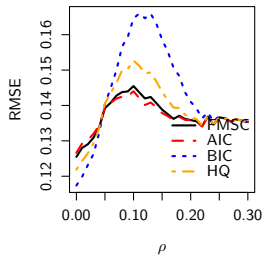
$N = 50, \gamma = 0.2$



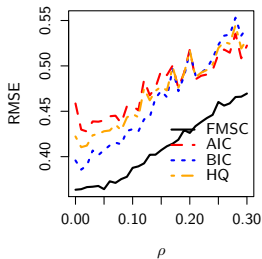
$N = 100, \gamma = 0.2$



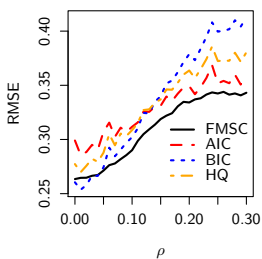
$N = 500, \gamma = 0.2$



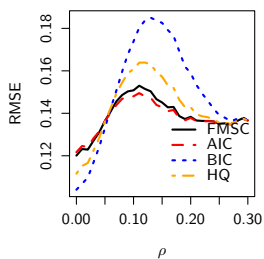
$N = 50, \gamma = 0.3$



$N = 100, \gamma = 0.3$



$N = 500, \gamma = 0.3$



# Empirical Example: Geography or Institutions?

## Institutions Rule

Acemoglu et al. (2001), Rodrik et al. (2004), Easterly & Levine (2003) – zero or negligible effects of “tropics, germs, and crops” in income per capita, controlling for institutions.

## Institutions *Don't* Rule

Sachs (2003) – Large negative direct effect of malaria transmission on income.

## Carstensen & Gundlach (2006)

How robust is Sachs's result?

# Carstensen & Gundlach (2006)

## Both Regressors Endogenous

$$\ln GDPC_i = \beta_1 + \beta_2 \cdot INSTITUTIONS_i + \beta_3 \cdot MALARIA_i + \epsilon_i$$

## Robustness

- ▶ Various measures of *INSTITUTIONS*, *MALARIA*
- ▶ Various instrument sets
- ▶  $\beta_3$  remains large, negative and significant.

## 2SLS for All Results That Follow

# Expand on Instrument Selection Exercise

## FMSC and Corrected Confidence Intervals

1. FMSC – which instruments to estimate effect of malaria?
2. Correct CIs for Instrument Selection – effect of malaria still negative and significant?

## Measures of *INSTITUTIONS* and *MALARIA*

- ▶ *rule* – Average governance indicator (Kaufmann, Kray and Mastruzzi; 2004)
- ▶ *malfal* – Proportion of population at risk of malaria transmission in 1994 (Sachs, 2001)

# Instrument Sets

## Baseline Instruments – Assumed Valid

- ▶ *Inmort* – Log settler mortality (per 1000), early 19th century
- ▶ *maleco* – Index of stability of malaria transmission

## Further Instrument Blocks

Climate *frost, humid, latitude*

Europe *eurfrac, engfrac*

Openness *coast, trade*

	$\mu = \text{malfal}$			$\mu = \text{rule}$		
	FMSC	posFMSC	$\hat{\mu}$	FMSC	posFMSC	$\hat{\mu}$
(1) Valid	3.0	3.0	-1.0	1.3	1.3	0.9
(2) Climate	3.1	3.1	-0.9	1.0	1.0	1.0
(3) Open	2.3	2.4	-1.1	1.2	1.2	0.8
(4) Eur	1.8	2.2	-1.1	0.5	0.7	0.9
(5) Climate, Eur	0.9	2.0	-1.0	0.3	0.6	0.9
(6) Climate, Open	1.9	2.3	-1.0	0.5	0.8	0.9
(7) Open, Eur	1.6	1.8	-1.2	0.8	0.8	0.8
(8) Full	0.5	1.7	-1.1	0.2	0.6	0.8
> 90% CI FMSC	(-1.6, -0.6)			(0.5, 1.2)		
> 90% CI posFMSC	(-1.6, -0.6)			(0.6, 1.3)		

# Lecture #7 – High-Dimensional Regression I

QR Decomposition

Singular Value Decomposition

Review of Principal Component Analysis (PCA)

Ridge Regression

Principal Components Regression



# QR Decomposition

## Result

Any  $n \times k$  matrix  $A$  with full column rank can be decomposed as  $A = QR$ , where  $R$  is an  $k \times k$  upper triangular matrix and  $Q$  is an  $n \times k$  matrix with orthonormal columns.

## Notes

- ▶ Columns of  $A$  are *orthogonalized* in  $Q$  via Gram-Schmidt.
- ▶ Since  $Q$  has orthogonal columns,  $Q'Q = I_k$ .
- ▶ It is *not* in general true that  $QQ' = I$ .
- ▶ If  $A$  is square, then  $Q^{-1} = Q'$ .

# Different Conventions for the QR Decomposition

## Thin aka Economical QR

$Q$  is an  $n \times k$  with orthonormal columns ( `qr_econ` in Armadillo).

## Thick QR

$Q$  is an  $n \times n$  *orthogonal* matrix.

## Relationship between Thick and Thin

Let  $A = QR$  be the “thick” QR and  $A = Q_1 R_1$  be the “thin” QR:

$$A = QR = Q \begin{bmatrix} R_1 \\ 0 \end{bmatrix} = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ 0 \end{bmatrix} = Q_1 R_1$$

My preferred convention is the thin QR...

# Least Squares via QR Decomposition

Let  $X = QR$

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'y = [(QR)'(QR)]^{-1}(QR)'y \\ &= [R'Q'QR]^{-1}R'Q'y = (R'R)^{-1}R'Qy \\ &= R^{-1}(R')^{-1}R'Q'y = R^{-1}Q'y\end{aligned}$$

In other words,  $\hat{\beta}$  solves  $R\beta = Q'y$ .

## Why Bother?

Much easier and faster to solve  $R\beta = Q'y$  than the normal equations  $(X'X)\beta = X'y$  since  $R$  is **upper triangular**.

## Back-Substitution to Solve $R\beta = Q'y$

The product  $Q'y$  is a vector, call it  $v$ , so the system is simply

$$\begin{bmatrix} r_{11} & r_{12} & r_{13} & \cdots & r_{1,n-1} & r_{1k} \\ 0 & r_{22} & r_{23} & \cdots & r_{2,n-1} & r_{2k} \\ 0 & 0 & r_{33} & \cdots & r_{3,n-1} & r_{3k} \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & r_{k-1,k-1} & r_{k-1,k} \\ 0 & 0 & \cdots & 0 & 0 & r_k \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_{k-1} \\ \beta_k \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_{k-1} \\ v_k \end{bmatrix}$$

$\beta_k = v_k/r_k \Rightarrow$  substitute this into  $\beta_{k-1}r_{k-1,k-1} + \beta_k r_{k-1,k} = v_{k-1}$   
to solve for  $\beta_{k-1}$ , and so on.

## Calculating the Least Squares Variance Matrix $\sigma^2(X'X)^{-1}$

- ▶ Since  $X = QR$ ,  $(X'X)^{-1} = R^{-1}(R^{-1})'$
- ▶ Easy to invert  $R$ : just apply **repeated** back-substitution:
  - ▶ Let  $A = R^{-1}$  and  $\mathbf{a}_j$  be the  $j$ th column of  $A$ .
  - ▶ Let  $\mathbf{e}_j$  be the  $j$ th standard basis vector.
  - ▶ Inverting  $R$  is equivalent to solving  $R\mathbf{a}_1 = \mathbf{e}_1$ , followed by  $R\mathbf{a}_2 = \mathbf{e}_2, \dots, R\mathbf{a}_k = \mathbf{e}_k$ .
- ▶ If you enclose a matrix in `trimatu()` or `trimatl()`, and request the inverse  $\Rightarrow$  Armadillo will carry out backward or forward substitution, respectively.

## QR Decomposition for Orthogonal Projections

Let  $X$  have full column rank and define  $P_X = X(X'X)^{-1}X'$

$$P_X = QR(R'R)^{-1}R'Q' = QRR^{-1}(R')^{-1}R'Q' = QQ'$$

It is *not* in general true that  $QQ' = I$  even though  $Q'Q = I$  since  $Q$  need not be square in the economical QR decomposition.

# The Singular Value Decomposition (SVD)

Any  $m \times n$  matrix  $A$  of arbitrary rank  $r$  can be written

$$A = UDV' = (\text{orthogonal})(\text{diagonal})(\text{orthogonal})$$

- ▶  $U = m \times m$  orthog. matrix whose cols contain e-vectors of  $AA'$
- ▶  $V = n \times n$  orthog. matrix whose cols contain e-vectors of  $A'A$
- ▶  $D = m \times n$  matrix whose first  $r$  main diagonal elements are the *singular values*  $d_1, \dots, d_r$ . All other elements are zero.
- ▶ The singular values  $d_1, \dots, d_r$  are the square roots of the non-zero eigenvalues of  $A'A$  and  $AA'$ .
- ▶ (E-values of  $A'A$  and  $AA'$  could be zero but not negative)

## SVD for Symmetric Matrices

If  $A$  is **symmetric** then  $A = Q\Lambda Q'$  where  $\Lambda$  is a diagonal matrix containing the e-values of  $A$  and  $Q$  is an orthonormal matrix whose columns are the corresponding e-vectors. Accordingly:

$$AA' = (Q\Lambda Q')(Q\Lambda Q')' = Q\Lambda Q'Q\Lambda Q' = Q\Lambda^2 Q'$$

and similarly

$$A'A = (Q\Lambda Q')'(Q\Lambda Q') = Q\Lambda Q'Q\Lambda Q' = Q\Lambda^2 Q'$$

using the fact that  $Q$  is orthogonal and  $\Lambda$  diagonal. Thus, when  $A$  is symmetric the SVD reduces to  $U = V = Q$  and  $D = \sqrt{\Lambda^2}$  so that *negative* eigenvalues become *positive* singular values.



# The Economical SVD

- ▶ Number of singular values is  $r = \text{Rank}(A) \leq \max\{m, n\}$
- ▶ Some cols of  $U$  or  $V$  multiplied by zeros in  $D$
- ▶ Economical SVD: only keep columns in  $U$  and  $V$  that are multiplied by non-zeros in  $D$  (Armadillo: `svd_econ`)
- ▶ Summation form:  $A = \sum_{i=1}^r d_i \mathbf{u}_i \mathbf{v}_i'$  where  $d_1 \leq d_2 \leq \dots \leq d_r$
- ▶ Matrix form: 
$$\underset{(n \times p)}{A} = \underset{(n \times r)}{U} \underset{(r \times r)}{D} \underset{(r \times p)}{V'}$$

In the economical SVD,  $U$  and  $V$  may no longer be square, so they are not orthogonal matrices but their *columns* are still orthonormal.

# Principal Component Analysis (PCA)

## Notation

Let  $\mathbf{x}$  be a  $p \times 1$  random vector with variance-covariance matrix  $\Sigma$ .

## Optimization Problem

$$\alpha_1 = \arg \max_{\alpha} \text{Var}(\alpha' \mathbf{x}) \quad \text{subject to} \quad \alpha' \alpha = 1$$

## First Principal Component

The linear combination  $\alpha_1' \mathbf{x}$  is the **first principal component** of  $\mathbf{x}$ .

It is the direction along with  $\mathbf{x}$  has **maximal variation**

# Solving for $\alpha_1$

## Lagrangian

$$\mathcal{L}(\alpha_1, \lambda) = \alpha' \Sigma \alpha - \lambda(\alpha' \alpha - 1)$$

## First Order Condition

$$2(\Sigma \alpha_1 - \lambda \alpha_1) = 0 \iff (\Sigma - \lambda I_p) \alpha_1 = 0 \iff \Sigma \alpha_1 = \lambda \alpha_1$$

## Variance of 1st PC

$\alpha_1$  is an e-vector of  $\Sigma$  but which one? Substituting,

$$\text{Var}(\alpha'_1 \mathbf{x}) = \alpha'_1 (\Sigma \alpha_1) = \lambda \alpha'_1 \alpha_1 = \lambda$$

## Solution

Var. of 1st PC equals  $\lambda$  and this is what we want to **maximize**, so

$\alpha_1$  is the e-vector corresponding to the largest e-value.

# Subsequent Principal Components

## Additional Constraint

Construct 2nd PC by solving the same problem as before with the additional constraint that  $\alpha'_2 \mathbf{x}$  is uncorrelated with  $\alpha'_1 \mathbf{x}$ .

## $j$ th Principal Component

The linear combination  $\alpha'_j \mathbf{x}$  where  $\alpha_j$  is the e-vector corresponding to the  $j$ th largest e-value of  $\Sigma$ .

# Sample PCA

## Notation

$X = (n \times p)$  **centered** data matrix – columns are mean zero.

## SVD

$$X = UDV', \text{ thus } X'X = VDU'UDV' = VD^2V'$$

## Sample Variance Matrix

$S = n^{-1}X'X$  has same e-vectors as  $X'X$  – the columns of  $V$ !

## Sample PCA

Let  $\mathbf{v}_j$  be the  $j$ th column of  $V$ . Then,

$\mathbf{v}_j$  = PC loadings for  $j$ th PC of  $S$

$\mathbf{v}_j' \mathbf{x}_i$  = PC score for individual/time period  $i$

# Sample PCA

## PC scores for $j$ th PC

$$\mathbf{z}_j = \begin{bmatrix} z_{j1} \\ \vdots \\ z_{jn} \end{bmatrix} = \begin{bmatrix} \mathbf{v}_j' \mathbf{x}_1 \\ \vdots \\ \mathbf{v}_j' \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1' \mathbf{v}_j \\ \vdots \\ \mathbf{x}_n' \mathbf{v}_j \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix} \mathbf{v}_j = X \mathbf{v}_j$$

## Getting PC Scores from SVD

Since  $X = UDV'$  and  $V'V = I$ ,  $XV = UD$ , i.e.

$$\begin{bmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_p \end{bmatrix} = \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_r \end{bmatrix} \begin{bmatrix} d_1 & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & d_r \end{bmatrix}$$

Hence we see that  $\mathbf{z}_j = d_j \mathbf{u}_j$

## Properties of PC Scores $\mathbf{z}_j$

Since  $X$  has been de-meaned:

$$\bar{z}_j = \frac{1}{n} \sum_{i=1}^n \mathbf{v}_j' \mathbf{x}_i = \mathbf{v}_j' \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \right) = \mathbf{v}_j' \mathbf{0} = 0$$

Hence, since  $X'X = VD^2V'$

$$\frac{1}{n} \sum_{i=1}^n (z_{ji} - \bar{z}_j)^2 = \frac{1}{n} \sum_{i=1}^n z_{ji}^2 = \frac{1}{n} \mathbf{z}_j' \mathbf{z}_j = \frac{1}{n} (X\mathbf{v}_j)' (X\mathbf{v}_j) = \mathbf{v}_j' S \mathbf{v}_j = d_j^2 / n$$

## Ridge Regression – OLS with an $L_2$ Penalty

$$\hat{\beta}_{Ridge} = \arg \min_{\beta} (\mathbf{y} - X\beta)'(\mathbf{y} - X\beta) + \lambda\beta'\beta$$

- ▶ Add a penalty for large coefficients
- ▶  $\lambda$  = non-negative constant we choose: strength of penalty
- ▶  $X$  and  $\mathbf{y}$  assumed to be **de-meaned** (don't penalize intercept)
- ▶ Unlike OLS, Ridge Regression is **not scale invariant**
  - ▶ In OLS if we replace  $\mathbf{x}_1$  with  $c\mathbf{x}_1$  then  $\beta_1$  becomes  $\beta_1/c$ .
  - ▶ The same is not true for ridge regression!
  - ▶ Typical to **standardize**  $X$  before carrying out ridge regression



## Alternative Formulation of Ridge Regression Problem

$$\hat{\beta}_{Ridge} = \arg \min_{\beta} (\mathbf{y} - X\beta)'(\mathbf{y} - X\beta) \quad \text{subject to} \quad \beta'\beta \leq t$$

- ▶ Ridge Regression is like least squares “on a budget.”
- ▶ Make one coefficient larger  $\Rightarrow$  must make another one smaller.
- ▶ One-to-one mapping from  $t$  to  $\lambda$  (data-dependent)

## Ridge as Bayesian Linear Regression

If we ignore the intercept, which is unpenalized), Ridge Regression gives the **posterior mode** from the Bayesian regression model:

$$\begin{aligned}y|X, \beta, \sigma^2 &\sim N(X\beta, \sigma^2 I_n) \\ \beta &\sim N(\mathbf{0}, \tau^2 I_p)\end{aligned}$$

where  $\sigma^2$  is assumed known and  $\lambda = \sigma^2/\tau^2$ . (In this example, the posterior is normal so the mode equals the mean)

# Explicit Solution to the Ridge Regression Problem

Objective Function:

$$\begin{aligned}Q(\beta) &= (\mathbf{y} - X\beta)'(\mathbf{y} - X\beta) + \lambda\beta'\beta \\&= \mathbf{y}'\mathbf{y} - \beta'X\mathbf{y} - \mathbf{y}'X\beta + \beta'X'X\beta + \lambda\beta'I_p\beta \\&= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'X\beta + \beta'(X'X + \lambda I_p)\beta\end{aligned}$$

Recall the following facts about matrix differentiation

$$\partial(\mathbf{a}'\mathbf{x})/\partial\mathbf{x} = \mathbf{a}, \quad \partial(\mathbf{x}'A\mathbf{x})/\partial\mathbf{x} = (A + A')\mathbf{x}$$

Thus, since  $(X'X + \lambda I_p)$  is symmetric,

$$\frac{\partial}{\partial\beta}Q(\beta) = -2X'\mathbf{y} + 2(X'X + \lambda I_p)\beta$$

# Explicit Solution to the Ridge Regression Problem

Previous Slide:

$$\frac{\partial}{\partial \beta} Q(\beta) = -2X'\mathbf{y} + 2(X'X + \lambda I_p)\beta$$

First order condition:

$$X'\mathbf{y} = (X'X + \lambda I_p)\beta$$

Hence,

$$\hat{\beta}_{Ridge} = (X'X + \lambda I_p)^{-1}X'\mathbf{y}$$

But is  $(X'X + \lambda I_p)$  guaranteed to be invertible?

## Ridge Regression via OLS with “Dummy Observations”

Ridge regression solution is identical to

$$\arg \min_{\beta} \left( \tilde{\mathbf{y}} - \tilde{X}\beta \right)' \left( \tilde{\mathbf{y}} - \tilde{X}\beta \right)$$

where

$$\tilde{\mathbf{y}} = \begin{bmatrix} \mathbf{y} \\ \mathbf{0}_p \end{bmatrix}, \quad \tilde{X} = \begin{bmatrix} X \\ \sqrt{\lambda} I_p \end{bmatrix}$$

since:

$$\begin{aligned} \left( \tilde{\mathbf{y}} - \tilde{X}\beta \right)' \left( \tilde{\mathbf{y}} - \tilde{X}\beta \right) &= \begin{bmatrix} (\mathbf{y} - X\beta)' & (-\sqrt{\lambda}\beta)' \end{bmatrix} \begin{bmatrix} (\mathbf{y} - X\beta) \\ -\sqrt{\lambda}\beta \end{bmatrix} \\ &= (\mathbf{y} - X\beta)'(\mathbf{y} - X\beta) + \lambda\beta'\beta \end{aligned}$$

## Ridge Regression Solution is Always Unique

Ridge solution is **always unique**, even if there are more regressors than observations! This follows from the preceding slide:

$$\hat{\beta}_{Ridge} = \arg \min_{\beta} \left( \tilde{\mathbf{y}} - \tilde{X}\beta \right)' \left( \tilde{\mathbf{y}} - \tilde{X}\beta \right)$$

$$\tilde{\mathbf{y}} = \begin{bmatrix} \mathbf{y} \\ \mathbf{0}_p \end{bmatrix}, \quad \tilde{X} = \begin{bmatrix} X \\ \sqrt{\lambda} I_p \end{bmatrix}$$

Columns of  $\sqrt{\lambda} I_p$  are linearly independent, so columns of  $\tilde{X}$  are also linearly independent, **regardless** of whether the same holds for the columns of  $X$ .

# Efficient Calculations for Ridge Regression

## QR Decomposition

Write Ridge as OLS with “dummy observations” with  $\tilde{X} = QR$  so

$$\hat{\beta}_{Ridge} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{\mathbf{y}} = R^{-1}Q'\tilde{\mathbf{y}}$$

which we can obtain by back-solving the system  $R\hat{\beta}_{Ridge} = Q'\tilde{\mathbf{y}}$ .

## Singular Value Decomposition

If  $p \gg n$ , it's much faster to use the SVD rather than the QR decomposition because the rank of  $X$  will be  $n$ . For implementation details, see Murphy (2012; Section 7.5.2).

# Comparing Ridge and OLS

## Assumption

Centered data matrix  $X_{(n \times p)}$  with rank  $p$  so OLS estimator is unique.

## Economical SVD

- ▶  $X_{(n \times p)} = U_{(n \times p)} D_{(p \times p)} V'_{(p \times p)}$  with  $U'U = V'V = I_p$ ,  $D$  diagonal
- ▶ Hence:  $X'X = (UDV')'(UDV') = VDU'UDV' = VD^2V'$
- ▶ Since  $V$  is square it is an orthogonal matrix:  $VV' = I_p$



## Comparing Ridge and OLS – The “Hat Matrix”

Using  $X = UDV'$  and the fact that  $V$  is orthogonal,

$$\begin{aligned}H(\lambda) &= X(X'X + \lambda I_p)^{-1}X' = UDV'(VD^2V + \lambda VV')^{-1}VDU' \\&= UDV'(VD^2V' + \lambda VV')^{-1}VDU' \\&= UDV'[V(D^2 + \lambda I_p)V']^{-1}VDU' \\&= UDV'(V')^{-1}(D^2 + \lambda I_p)^{-1}(V)^{-1}VDU' \\&= UDV'V(D^2 + \lambda I_p)^{-1}V'VDU' \\&= UD(D^2 + \lambda I_p)^{-1}DU'\end{aligned}$$

# Model Complexity of Ridge Versus OLS

## OLS Case

Number of free parameters equals number of parameters  $p$ .

## Ridge is more complicated

Even though there are  $p$  parameters they are **constrained!**

Idea: use trace of  $H(\lambda)$

$$\text{df}(\lambda) = \text{tr} \{H(\lambda)\} = \text{tr} \{X(X'X + \lambda I_p)^{-1}X'\}$$

Why? Works for OLS:  $\lambda = 0$

$$\text{df}(0) = \text{tr} \{H(0)\} = \text{tr} \{X(X'X)^{-1}X'\} = p$$

# Effective Degrees of Freedom for Ridge Regression

Using cyclic permutation property of trace:

$$\begin{aligned}\text{df}(\lambda) &= \text{tr} \{H(\lambda)\} = \text{tr} \{X(X'X + \lambda I_p)^{-1}X'\} \\&= \text{tr} \{UD (D^2 + \lambda I_p)^{-1} DU'\} \\&= \text{tr} \{DU'UD (D^2 + \lambda I_p)^{-1}\} \\&= \text{tr} \{D^2 (D^2 + \lambda I_p)^{-1}\} \\&= \sum_{j=1}^p \frac{d_j^2}{d_j^2 + \lambda}\end{aligned}$$

- ▶  $\text{df}(\lambda) \rightarrow 0$  as  $\lambda \rightarrow \infty$
- ▶  $\text{df}(\lambda) = p$  when  $\lambda = 0$
- ▶  $\text{df}(\lambda) < p$  when  $\lambda > 0$

## Comparing OLS and Ridge Predictions

$$\begin{aligned}\hat{y}(\lambda) &= X\hat{\beta}(\lambda) = X(X'X + \lambda I_p)^{-1}X'y \\ &= H(\lambda)y = \left[UD(D^2 + \lambda I_p)^{-1}DU'\right]y \\ &= \left[\sum_{j=1}^p \mathbf{u}_j \left(\frac{d_j^2}{d_j^2 + \lambda}\right) \mathbf{u}_j'\right]y = \sum_{j=1}^p \left(\frac{d_j^2}{d_j^2 + \lambda}\right) \mathbf{u}_j \mathbf{u}_j' y\end{aligned}$$

## Comparing OLS and Ridge Predictions

$$\hat{y}(\lambda) = \sum_{j=1}^p \left( \frac{d_j^2}{d_j^2 + \lambda} \right) \mathbf{u}_j \mathbf{u}_j' \mathbf{y}$$

- ▶ Since  $X$  is centered,  $\mathbf{z}_j = d_j \mathbf{u}_j$  is the  $j$ th sample PC
- ▶  $d_j^2$  is proportional to the **variance** of the  $j$ th sample PC
- ▶ Prediction from regression of  $\mathbf{y}$  on  $\mathbf{z}_j$  is:

$$\mathbf{z}_j (\mathbf{z}_j' \mathbf{z}_j)^{-1} \mathbf{z}_j' \mathbf{y} = d_j \mathbf{u}_j (d_j^2 \mathbf{u}_j' \mathbf{u}_j)^{-1} d_j \mathbf{u}_j' \mathbf{y} = \mathbf{u}_j \mathbf{u}_j' \mathbf{y}$$

- ▶ Ridge equivalent to regressing  $y$  on sample PCs of  $X$  but shrinking predictions to zero: higher variance PCs are shrunk less.
- ▶ OLS doesn't shrink.

# Principal Components Regression (PCR)

Instead of “smooth weights” as in Ridge, truncate the PCs:

1. Calculate SVD  $X = UDV'$  of **centered** data matrix  $X$
2. Construct the sample principal components:  $\mathbf{z}_j = d_j \mathbf{u}_j$ .
3. Throw away all but first  $k$  principal components, where  $k < p$ .
4. Regress  $\mathbf{y}$  on  $\mathbf{z}_1, \dots, \mathbf{z}_M$ .

## PCR versus Ridge

- ▶ PCR is a much less smooth version of Ridge
- ▶ Conventional wisdom is that PCR will perform worse since it shrinks low variance directions too much and doesn't shrink high variance directions at all.
- ▶ However, Dhillon et al. (2013) show that the MSE risk of PCR is always within a constant factor of that of Ridge Regression while there are situations in which Ridge can be arbitrarily worse than PCR in terms of MSE.
- ▶ In practice, which is better depends on the DGP

# Lecture #8 – High-Dimensional Regression II

LASSO

Overview of Factor Models

Choosing the Number of Factors

Diffusion Index Forecasting



# Least Absolute Shrinkage and Selection Operator (LASSO)

Bühlmann & van de Geer (2011); Hastie, Tibshirani & Wainwright (2015)

Assume that  $X$  has been centered: don't penalize intercept!

## Notation

$$\|\beta\|_2^2 = \sum_{j=1}^p \beta_j^2, \quad \|\beta\|_1 = \sum_{j=1}^p |\beta_j|$$

## Ridge Regression – $L_2$ Penalty

$$\hat{\beta}_{Ridge} = \arg \min_{\beta} (\mathbf{y} - X\beta)'(\mathbf{y} - X\beta) + \lambda \|\beta\|_2^2$$

## LASSO – $L_1$ Penalty

$$\hat{\beta}_{Lasso} = \arg \min_{\beta} (\mathbf{y} - X\beta)'(\mathbf{y} - X\beta) + \lambda \|\beta\|_1$$

# Other Ways of Thinking about LASSO

## Constrained Optimization

$$\arg \min_{\beta} (\mathbf{y} - X\beta)'(\mathbf{y} - X\beta) \quad \text{subject to} \quad \sum_{j=1}^p |\beta_j| \leq t$$

Data-dependent, one-to-one mapping between  $\lambda$  and  $t$ .

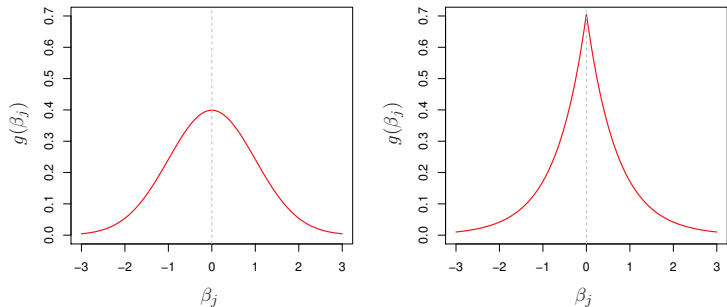
## Bayesian Posterior Mode

Ignoring the intercept, LASSO is the posterior model for  $\beta$  under

$$\mathbf{y}|X, \beta, \sigma^2 \sim N(X\beta, \sigma^2 I_n), \quad \beta \sim \prod_{j=1}^p \text{Lap}(\beta_j|0, \tau)$$

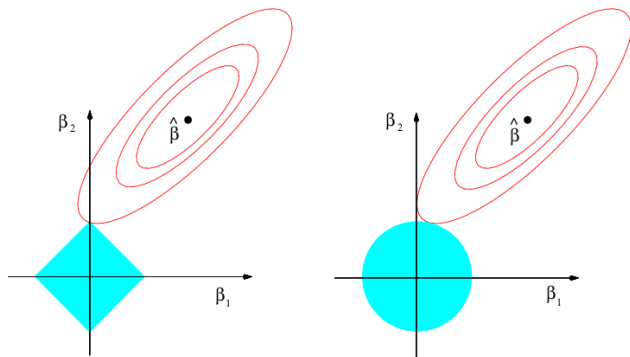
where  $\lambda = 1/\tau$  and  $\text{Lap}(x|\mu, \tau) = (2\tau)^{-1} \exp \{-\tau^{-1}|x - \mu|\}$

# Comparing Ridge and LASSO – Bayesian Posterior Modes



**Figure:** Ridge, at left, puts a normal prior on  $\beta$  while LASSO, at right, uses a Laplace prior, which has fatter tails and a taller peak at zero.

## Comparing LASSO and Ridge – Constrained OLS



**Figure:**  $\hat{\beta}$  denotes the MLE and the ellipses are the contours of the likelihood. LASSO, at left, and Ridge, at right, both shrink  $\beta$  away from the MLE towards zero. Because of its diamond-shaped constraint set, however, LASSO favors a **sparse solution** while Ridge does not

# No Closed-Form for LASSO!

## Simple Special Case

Suppose that  $X'X = I_p$

## Maximum Likelihood

$$\hat{\beta}_{MLE} = (X'X)^{-1}X'y = X'y, \quad \hat{\beta}_j^{MLE} = \sum_{i=1}^n x_{ij}y_i$$

## Ridge Regression

$$\hat{\beta}_{Ridge} = (X'X + \lambda I_p)^{-1}X'y = [(1 + \lambda)I_p]^{-1}\hat{\beta}_{MLE}, \quad \hat{\beta}_j^{Ridge} = \frac{\hat{\beta}_j^{MLE}}{1 + \lambda}$$

So what about LASSO?

LASSO when  $X'X = I_p$  so  $\hat{\beta}_{MLE} = X'y$

Want to Solve

$$\hat{\beta}_{LASSO} = \arg \min_{\beta} (\mathbf{y} - X\beta)'(\mathbf{y} - X\beta) + \lambda \|\beta\|_1$$

Expand First Term

$$\begin{aligned}(\mathbf{y} - X\beta)'(\mathbf{y} - X\beta) &= \mathbf{y}'\mathbf{y} - 2\beta'X'\mathbf{y} + \beta'X'X\beta \\ &= (\text{constant}) - 2\beta'\hat{\beta}_{MLE} + \beta'\beta\end{aligned}$$

Hence

$$\begin{aligned}\hat{\beta}_{LASSO} &= \arg \min_{\beta} (\beta'\beta - 2\beta'\hat{\beta}_{MLE}) + \lambda \|\beta\|_1 \\ &= \arg \min_{\beta} \sum_{j=1}^p \left( \beta_j^2 - 2\beta_j\hat{\beta}_j^{MLE} + \lambda |\beta_j| \right)\end{aligned}$$

# LASSO when $X'X = I_p$

## Preceding Slide

$$\hat{\beta}_{LASSO} = \arg \min_{\beta} \sum_{j=1}^p \left( \beta_j^2 - 2\beta_j \hat{\beta}_j^{MLE} + \lambda |\beta_j| \right)$$

## Key Simplification

Equivalent to solving  $j$  independent optimization problems:

$$\hat{\beta}_j^{Lasso} = \arg \min_{\beta_j} \left( \beta_j^2 - 2\beta_j \hat{\beta}_j^{MLE} + \lambda |\beta_j| \right)$$

- ▶ Sign of  $\beta_j^2$  and  $\lambda |\beta_j|$  unaffected by  $\text{sign}(\beta_j)$
- ▶  $\hat{\beta}_j^{MLE}$  is a function of data only – outside our control
- ▶ Minimization requires **matching**  $\text{sign}(\beta_j)$  to  $\text{sign}(\hat{\beta}_j^{MLE})$

## LASSO when $X'X = I_p$

Case I:  $\hat{\beta}^{MLE} > 0 \implies \beta_j > 0 \implies |\beta_j| = \beta_j$

Optimization problem becomes

$$\hat{\beta}_j^{Lasso} = \arg \min_{\beta_j} \beta_j^2 - 2\beta_j \hat{\beta}_j^{MLE} + \lambda \beta_j$$

Interior solution:

$$\hat{\beta}_j = \hat{\beta}_j^{MLE} - \frac{\lambda}{2}$$

Can't have  $\beta_j < 0$ : corner solution sets  $\beta_j = 0$

$$\hat{\beta}_j^{Lasso} = \max \left\{ 0, \hat{\beta}_j^{MLE} - \frac{\lambda}{2} \right\}$$



## LASSO when $X'X = I_p$

Case II:  $\hat{\beta}^{MLE} \leq 0 \implies \beta_j \leq 0 \implies |\beta_j| = -\beta_j$

Optimization problem becomes

$$\hat{\beta}_j^{Lasso} = \arg \min_{\beta_j} \beta_j^2 - 2\beta_j \hat{\beta}_j^{MLE} - \lambda \beta_j$$

Interior solution:

$$\hat{\beta}_j = \hat{\beta}_j^{MLE} + \frac{\lambda}{2}$$

Can't have  $\beta_j > 0$ : corner solution sets  $\beta_j = 0$

$$\hat{\beta}_j^{Lasso} = \min \left\{ 0, \hat{\beta}_j^{MLE} + \frac{\lambda}{2} \right\}$$

## Ridge versus LASSO when $X'X = I_p$

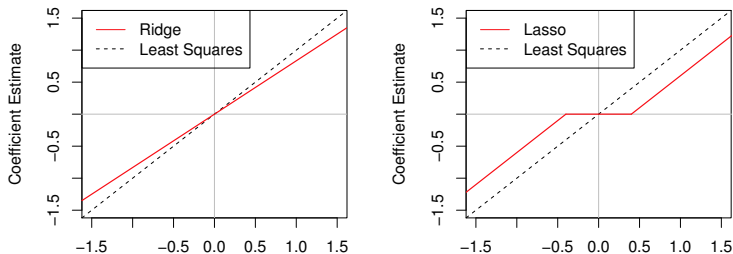


Figure: Horizontal axis in each plot is MLE

$$\hat{\beta}_j^{Ridge} = \left( \frac{1}{1 + \lambda} \right) \hat{\beta}_j^{MLE}$$

$$\hat{\beta}_j^{Lasso} = \text{sign} \left( \hat{\beta}_j^{MLE} \right) \max \left\{ 0, \left| \hat{\beta}_j^{MLE} \right| - \frac{\lambda}{2} \right\}$$

# Calculating LASSO – The Shooting Algorithm

## Cyclic Coordinate Descent

**Data:**  $\mathbf{y}$ ,  $X$ ,  $\lambda \geq 0$ ,  $\varepsilon > 0$

**Result:** LASSO Solution

$\beta \leftarrow \text{ridge}(X, \mathbf{y}, \lambda)$

**repeat**

$\beta^{\text{prev}} \leftarrow \beta$

**for**  $j = 1, \dots, p$  **do**

$a_j \leftarrow 2 \sum_{i=1}^n x_{ij}^2$

$c_j \leftarrow 2 \sum_{i=1}^n x_{ij}(y_i - \mathbf{x}_i' \beta + \beta_j x_{ij})$

$\beta_j \leftarrow \text{sign}(c_j/a_j) \max \{0, |c_j/a_j| - \lambda/a_j\}$

**end**

**until**  $\sum_{j=1}^p |\beta_j^{\text{prev}} - \beta_j| < \varepsilon;$

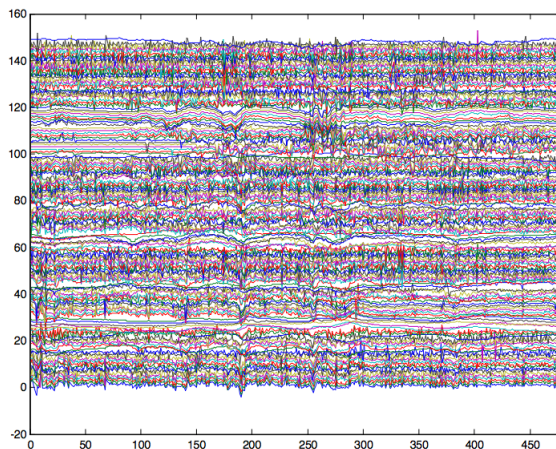
# The Basic Idea

- ▶  $(T \times N)$  Matrix  $X$  of observations
- ▶  $X_t$  contains a large number  $N$  of time series
- ▶ Comparable number  $T$  of time periods
- ▶ Can we “summarize” this information in some useful way?
- ▶ Forecasting and policy analysis applications

## Survey Articles

Stock & Watson (2010), Bai & Ng (2008), Stock & Watson (2006)

## Example: Stock and Watson Dataset



Monthly Macroeconomic Indicators:  $N > 200$ ,  $T > 400$

# Classical Factor Analysis Model

Assume that  $X_t$  has been de-meanned...

$$\underset{(N \times 1)}{X_t} = \underset{(r \times 1)}{\Lambda} F_t + \epsilon_t$$

$$\begin{bmatrix} F_t \\ \epsilon_t \end{bmatrix} \stackrel{iid}{\sim} \mathcal{N} \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} I_r & 0 \\ 0 & \Psi \end{bmatrix} \right)$$

$\Lambda$  = matrix of factor loadings

$\Psi$  = diagonal matrix of idiosyncratic variances.

## Adding Time-Dependence

$$\underset{(N \times 1)}{X_t} = \Lambda \underset{(r \times 1)}{F_t} + \epsilon_t$$

$$\underset{(r \times 1)}{F_t} = A_1 F_{t-1} + \dots + A_p F_{t-p} + u_t$$

$$\begin{bmatrix} u_t \\ \epsilon_t \end{bmatrix} \stackrel{iid}{\sim} \mathcal{N} \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} I_r & 0 \\ 0 & \Psi \end{bmatrix} \right)$$

# Terminology

**Static**  $X_t$  depends only on  $F_t$

**Dynamic**  $X_t$  depends on lags of  $F_t$  as well

**Exact**  $\Psi$  is diagonal and  $\epsilon_t$  independent over time

**Approximate** Some cross-sectional & temporal dependence in  $\epsilon_t$

The model I wrote down on the previous slide is sometimes called an “exact, static factor model” even though  $F_t$  has dynamics.



# Some Caveats

1. Are “static” and “dynamic” really different?
  - ▶ Can write dynamic model as a static one with more factors
  - ▶ Static representation involves “different” factors, but we may not care: are the factors “real” or just a data summary?
2. Can we *really* allow for cross-sectional dependence?
  - ▶ Unless the off-diagonal elements of  $\Psi$  are close to zero we can't tell them apart from the common factors
  - ▶ “Approximate” factor models basically assume conditions under which the off-diagonal elements of  $\Psi$  are negligible
  - ▶ Similarly, time series dependence in  $\epsilon_t$  can't be very strong (stationary ARMA is ok)

# Methods of Estimation for Dynamic Factor Models

1. Bayesian Estimation
2. Maximum Likelihood: EM-Algorithm + Kalman Filter
  - ▶ Watson & Engle (1983); Ghahramani & Hinton (1996); Jungbacker & Koopman (2008); Doz, Giannone & Reichlin (2012)
3. “Nonparametric” Estimation via PCA
  - ▶ PCA on the  $(T \times N)$  matrix  $X$ , ignoring time dependence.
  - ▶ The  $(r \times 1)$  vector  $\hat{F}_t$  of PC scores associated with the first  $r$  PCs are our estimate of  $F_t$
  - ▶ Essentially treats  $F_t$  as an  $r$ -dimensional *parameter* to be estimated from an  $N$ -dimensional observation  $X_t$

# Estimation by PCA

## PCA Normalization

- ▶  $F'F/T = I_r$  where  $F = (F_1, \dots, F_T)'$
- ▶  $\Lambda'\Lambda = \text{diag}(\mu_1, \dots, \mu_r)$  where  $\mu_1 \geq \mu_2 \geq \dots \geq \mu_r$

## Assumption I

Factors are *pervasive*:  $\Lambda'\Lambda/N \rightarrow D_\Lambda$  an  $(r \times r)$  full rank matrix.

## Assumption II

max e-value  $E[\epsilon_t \epsilon_t'] \leq c \leq \infty$  for all  $N$ .

## Upshot of the Assumptions

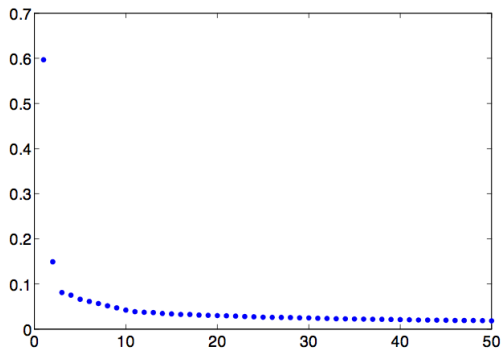
Average over the cross-section  $\implies$  contribution from the factors persists while contribution from the idiosyncratic terms disappears as  $N \rightarrow \infty$ .

## Key Result for PCA Estimation

Under the assumptions on the previous slide and some other technical conditions, the first  $r$  PCs of  $X$  consistently estimate the space spanned by the factors as  $N, T \rightarrow \infty$ .

## Choosing the Number of Factors – Scree Plot

If we use PC estimation, we can look at something called a “scree plot” to help us decide how many PCs to include:



This figure depicts the eigenvalues for an  $N = 1148$ ,  $T = 252$  dataset of excess stock returns

## Choosing the Number of Factors – Bai & Ng (2002)

Choose  $r$  to minimize an information criterion:

$$IC(r) = \log V_r(\hat{\Lambda}, \hat{F}) + r \cdot g(N, T)$$

where

$$V_r(\Lambda, F) = \frac{1}{NT} \sum_{t=1}^T (X_t - \Lambda F_t)' (X_t - \Lambda F_t)$$

and  $g$  is a penalty function. The paper provides conditions on the penalty function that guarantee consistent estimation of the “true number” of factors.

# Some Special Problems in High-dimensional Forecasting

## Estimation Uncertainty

We've already seen that OLS can perform very badly if the number of regressors is large relative to sample size.

## Best Subsets Infeasible

With more than 30 or so regressors, we can't check all subsets of predictors making classical model selection problematic.

## Noise Accumulation

Large  $N$  is supposed to help in factor models: averaging over the cross-section gives a consistent estimator of factor space. This can fail in practice, however, since it relies on the assumption that the factors are *pervasive*. See Boivin & Ng (2006).

# Diffusion Index Forecasting – Stock & Watson (2002a,b)

JASA paper has the theory, JBES paper has macro forecasting example.

## Basic Setup

Forecast scalar time series  $y_{t+1}$  using  $N$ -dimensional collection of time series  $X_t$  where we observe periods  $t = 1, \dots, T$ .

## Assumption

Static representation of Dynamic Factor Model:

$$y_t = \beta' F_t + \gamma(L)y_t + \epsilon_{t+1}$$

$$X_t = \Lambda F_t + e_t$$

## “Direct” Multistep Ahead Forecasts

“Iterated” forecast would be linear in  $F_t$ ,  $y_t$  and lags:

$$y_{t+h}^h = \alpha_h + \beta_h(L)F_t + \gamma_h(L)y_t + \epsilon_{t+h}^h$$



This is really just PCR

# Diffusion Index Forecasting – Stock & Watson (2002a,b)

## Estimation Procedure

### 1. Data Pre-processing

- 1.1 Transform all series to stationarity (logs or first difference)
- 1.2 Center and standardize all series
- 1.3 Remove outliers (ten times IQR from median)
- 1.4 Optionally augment  $X_t$  with lags

### 2. Estimate the Factors

- ▶ No missing observations: PCA on  $X_t$  to estimate  $\hat{F}_t$
- ▶ Missing observations/Mixed-frequency: EM-algorithm

### 3. Fit the Forecasting Regression

- ▶ Regress  $y_t$  on a constant and lags of  $\hat{F}_t$  and  $y_t$  to estimate the parameters of the “Direct” multistep forecasting regression.

# Diffusion Index Forecasting – Stock & Watson (2002b)

Recall from above that, under certain assumptions, PCA consistently estimates the space spanned by the factors. Broadly similar assumptions are at work here.

## Main Theoretical Result

Moment restrictions on  $(\epsilon, e, F)$  plus a “rank condition” on  $\Lambda$  imply that the MSE of the procedure on the previous slide converges to that of the infeasible optimal procedure, provided that  $N, T \rightarrow \infty$ .

# Diffusion Index Forecasting – Stock & Watson (2002a)

## Forecasting Experiment

- ▶ Simulated real-time forecasting of eight monthly macro variables from 1959:1 to 1998:12
- ▶ Forecasting Horizons: 6, 12, and 24 months
- ▶ “Training Period” 1959:1 through 1970:1
- ▶ Predict  $h$ -steps ahead out-of-sample, roll and re-estimate.
- ▶ BIC to select lags and # of Factors in forecasting regression
- ▶ Compare Diffusion Index Forecasts to Benchmark
  - ▶ AR only
  - ▶ Factors only
  - ▶ AR + Factors

# Diffusion Index Forecasting – Stock & Watson (2002a)

## Empirical Results

- ▶ Factors provide a substantial improvement over benchmark forecasts in terms of MSPE
- ▶ Six factors explain 39% of the variance in the 215 series; twelve explain 53%
- ▶ Using all 215 series tends to work better than restricting to balanced panel of 149 (PCA estimation)
- ▶ Augmenting  $X_t$  with lags isn't helpful

# Lecture #10 – Selective Inference

Optimal Inference After Model Selection (Fithian et al., 2017)

# How Statistics is Done In Reality

## Step 1: Selection – Decide what questions to ask.

“The analyst chooses a statistical model for the data at hand, and formulates testing, estimation, or other problems in terms of unknown aspects of that model.”

## Step 2: Inference – Answer the Questions.

“The analyst investigates the chosen problems using the data and the selected model.”

## Problem – “Data-snooping”

Standard techniques for (frequentist) statistical inference assume that we choose our questions **before** observing the data.

## Simple Example: “File Drawer Problem”

$Y_i \sim \text{iid } N(\mu_i, 1)$  for  $i = 1, \dots, n$

- ▶ I want to know which  $\mu_i \neq 0$ , but I'm busy and  $n$  is big.
- ▶ My RA looks at each  $Y_i$  and finds the “interesting” ones, namely  $\hat{\mathcal{I}} = \{i: |Y_i| > 1\}$ .
- ▶ I test  $H_{0,i}: \mu_i = 0$  against the two-sided alternative at the 5% significance level for each  $i \in \hat{\mathcal{I}}$ .

### Two Questions

1. What is the probability of falsely rejecting  $H_{0,i}$ ?
2. Among all  $H_{0,i}$  that I test, what fraction are false rejections?



## Simple Example: “File Drawer Problem”

$$\begin{aligned}\mathbb{P}_{H_{0,i}}(\{\text{Reject } H_{0,i}\}) &= \mathbb{P}_{H_{0,i}}(\{\text{Test } H_{0,i}\} \cap \{\text{Reject } H_{0,i}\}) \\&= \mathbb{P}_{H_{0,i}}(\{\text{Reject } H_{0,i}\} | \{\text{Test } H_{0,i}\}) \mathbb{P}_{H_{0,i}}(\{\text{Test } H_{0,i}\}) \\&= \mathbb{P}_{H_{0,i}}(|Y_i| > 1.96 | |Y_i| > 1) \mathbb{P}_{H_{0,i}}(|Y_i| > 1) \\&= \frac{2\Phi(-1.96)}{2\Phi(-1)} \times 2\Phi(-1) \\&\approx 0.16 \times 0.32 \approx 0.05\end{aligned}$$

$$\begin{aligned}\mathbb{P}_{H_{0,i}}(\{\text{Reject } H_{0,i}\} | \{\text{Test } H_{0,i}\}) &= \mathbb{P}_{H_{0,i}}(|Y_i| > 1.96 | |Y_i| > 1) \\&= \frac{\Phi(-1.96)}{\Phi(-1)} \approx 0.16\end{aligned}$$

# Simple Example: “File Drawer Problem”

## Conditional vs. Unconditional Type I Error Rates

- ▶ The **conditional** probability of falsely rejecting  $H_{0,i}$ , given that I have tested it, is about 0.16.
- ▶ The **unconditional** probability of falsely rejecting  $H_{0,i}$  is 0.05 since I only test a false null with probability 0.32.

## Idea for Post-Selection Inference

Control the Type I Error Rate **conditional on selection**: “The answer must be valid, given that the question was asked.”

## Simple Example: “File Drawer Problem”

### Conditional Type I Error Rate

Solve  $\mathbb{P}_{H_{0,i}}(\{|Y_i| > c\}|\{|Y_i| > 1\}) = 0.05$  for  $c$ .

$$\mathbb{P}_{H_{0,i}}(\{|Y_i| > c\}|\{|Y_i| > 1\}) = \frac{\Phi(-c)}{\Phi(-1)} = 0.05$$

$$c = -\Phi^{-1}(\Phi(-1) \times 0.05)$$

$$c \approx 2.41$$

### Notice:

To account for the first-stage selection step, we need a larger critical value: 2.41 vs. 1.96. This means the test is less powerful.

# Selective Inference vs. Sample-Splitting

## Classical Inference

Control the Type I error under model  $M$ :  $\mathbb{P}_{M,H_0}(\text{reject } H_0) \leq \alpha$ .

## Selective Inference

Control the Type I error under model  $M$ , **given** that  $M$  and  $H_0$  were selected:  $\mathbb{P}_{M,H_0}(\text{reject } H_0 | \{M, H_0 \text{ selected}\}) \leq \alpha$ .

## Sample-Splitting

Use different datasets to choose  $(M, H_0)$  and carry out inference:

$$\mathbb{P}_{M,H_0}(\text{reject } H_0 | \{M, H_0 \text{ selected}\}) = \mathbb{P}_{M,H_0}(\text{reject } H_0).$$

# Selective Inference in Exponential Family Models

## Questions

1. Recipe for selective inference in realistic examples?
2. How to construct the “best” selective test in a given example?
3. How does selective inference compare to sample-splitting?

## Fithian, Sun & Taylor (2017)

- ▶ Use classical theory for exponential family models (Lehmann & Scheffé).
- ▶ Computational procedure for UMPU selective test/CI after arbitrary model/hypothesis selection.
- ▶ Sample-splitting is typically inadmissible (wastes information).
- ▶ Example: post-selection inference for high-dimensional regression

# A Prototype Example of Selective Inference

This is my own example, but uses the same idea that underlies Fithian et al.

- ▶ Choose between two models on a parameter  $\delta$ .
  - ▶ If  $\delta \neq 0$ , choose M1; if  $\delta = 0$ , choose M2
  - ▶ E.g.  $\delta$  is the endogeneity of  $X$ , M1 is IV and M2 is OLS
- ▶ Observe  $Y_\delta \sim N(\delta, \sigma_\delta^2)$  and use this to choose a model.
  - ▶ Selection Event:  $A \equiv \{|Y_\delta| > c\}$ , for some critical value  $c$
  - ▶ If  $A$ , then choose M1. Otherwise, choose M2.
- ▶ After choosing a model, carry out inference for  $\beta$ .
  - ▶ Under a particular model  $M$ ,  $Y_\beta \sim N(\beta, \sigma_\beta^2)$
  - ▶  $\beta$  is a *model-specific* parameter: could be meaningless or not even exist under a different model.
- ▶ If  $Y_\beta$  and  $Y_\delta$  are correlated (under model M), we need to account for conditioning on  $A$  when carrying out inference for  $\beta$ .

# All Calculations are Under a Given Model $M$

## Key Idea

Under whichever model  $M$  ends up being selected, there is a joint normal distribution for  $Y_\beta$  and  $Y_\delta$  *without* conditioning on  $A$ .

WLOG unit variances,  $\rho$  known

$$\begin{bmatrix} Y_\beta \\ Y_\delta \end{bmatrix} \sim N \left( \begin{bmatrix} \beta \\ \delta \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)$$

As long as we can consistently estimate the variances of  $Y_\beta$  and  $Y_\delta$  along with their covariance, this is not a problem.

# Selective Inference in a Bivariate Normal Example

$$\begin{bmatrix} Y_\beta \\ Y_\delta \end{bmatrix} \sim N \left( \begin{bmatrix} \beta \\ \delta \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right), \quad A \equiv \{|Y_\delta| > c\}$$

## Two Cases

1. Condition on  $A$  occurring
2. Condition on  $A$  *not* occurring

## Problem

If  $\delta$  were known, we could directly calculate how conditioning on  $A$  affects the distribution of  $Y_\beta$ , but  $\delta$  is unknown!

## Solution

Condition on a sufficient statistic for  $\delta$ .



# Conditioning on a Sufficient Statistic

## Theorem

If  $U$  is a sufficient statistic for  $\delta$ , then the joint distribution of  $(Y_\beta, Y_\delta)$  given  $U$  does not depend on  $\delta$ .

## In Our Example

Residual  $U = Y_\delta - \rho Y_\beta$  from a projection of  $Y_\delta$  onto  $Y_\beta$  is sufficient for  $\delta$ .

## Straightforward Calculation

$$\begin{bmatrix} Y_\beta \\ Y_\delta \end{bmatrix} \bigg| (U = u) = \begin{bmatrix} \beta + Z \\ u + \rho(\beta + Z) \end{bmatrix}, \quad Z \sim N(0, 1)$$

Notice that this is a singular normal distribution

## The Distribution of $Y_\beta|(A, U = u)$

$$\begin{bmatrix} Y_\beta \\ Y_\delta \end{bmatrix} \bigg| (U = u) = \begin{bmatrix} \beta + Z \\ u + \rho(\beta + Z) \end{bmatrix}, \quad Z \sim N(0, 1)$$

Start with case in which  $A$  occurs so we select  $M1$ . Under  $H_0: \beta = \beta_0$ ,

$$\begin{aligned} \mathbb{P}_{\beta_0}(Y_\beta \leq y | A, U = u) &= \frac{\mathbb{P}_{\beta_0}(\{Y_\beta \leq y\} \cap A | U = u)}{\mathbb{P}_{\beta_0}(A | U = u)} \\ &= \frac{\mathbb{P}(\{Z \leq y - \beta_0\} \cap \{|u + \rho(\beta_0 + Z)| > c\})}{\mathbb{P}(|u + \rho(\beta_0 + Z)| > c)} \end{aligned}$$

$\mathbb{P}(A|U = u)$  under  $H_0: \beta = \beta_0$

$$\begin{aligned}P_D(A) &\equiv P_{\beta_0}(A|U = u) \\&= \mathbb{P}(|u + \rho(\beta_0 + Z)| > c) \\&= \mathbb{P}[u + \rho(\beta_0 + Z) > c] + \mathbb{P}[u + \rho(\beta_0 + Z) < -c] \\&= \mathbb{P}[\rho(\beta_0 + Z) > c - u] + \mathbb{P}[u + \rho(\beta_0 + Z) < -c - u] \\&= 1 - \Phi\left(\frac{c - u}{\rho} - \beta_0\right) + \Phi\left(\frac{-c - u}{\rho} - \beta_0\right)\end{aligned}$$

$\mathbb{P}(\{Y_\beta \leq y\} \cap A | U = u)$  under  $H_0: \beta = \beta_0$

$$\begin{aligned} P_N(A) &\equiv \mathbb{P}(\{Y_\beta \leq y\} \cap A | U = u) \\ &= \mathbb{P}(\{Z \leq y - \beta_0\} \cap \{|u + \rho(\beta_0 + Z)| > c\}) \\ &= \begin{cases} \Phi(y - \beta_0), & y < (-c - u)/\rho \\ \Phi\left(\frac{-c - u}{\rho} - \beta_0\right), & (-c - u)/\rho \leq y \leq (c - u)/\rho \\ \Phi(y - \beta_0) - \Phi\left(\frac{c - u}{\rho} - \beta_0\right) + \Phi\left(\frac{-c - u}{\rho} - \beta_0\right), & y > (c - u)/\rho \end{cases} \end{aligned}$$

$$F_{\beta_0}(y|A, U = u)$$

Define  $\ell(u) = (-c - u)/\rho$ ,  $r(u) = (c - u)/\rho$ . We have:

$$F_{\beta_0}(y|A, U = u) = P_N(A)/P_D(A)$$

where

$$P_D(A) \equiv 1 - \Phi(r(u) - \beta_0) + \Phi(\ell(u) - \beta_0)$$

$$P_N(A) \equiv \begin{cases} \Phi(y - \beta_0), & y < \ell(u) \\ \Phi(\ell(u) - \beta_0), & \ell(u) \leq y \leq r(u) \\ \Phi(y - \beta_0) - \Phi(r(u) - \beta_0) + \Phi(\ell(u) - \beta_0), & y > r(u) \end{cases}$$

Note that  $F_{\beta_0}(y|A, U = u)$  has a *flat region* where  $\ell(u) \leq y \leq r(u)$

$$Q_{\beta_0}(p|A, U = u)$$

Inverting the CDF from the preceding slide:

$$Q_{\beta_0}(p|A, U = u) = \begin{cases} \beta_0 + \Phi^{-1}(p \times P_D(A)), & p < p^* \\ \beta_0 + \Phi^{-1}[p \times P_D(A) + \Phi(r(u) - \beta_0) - \Phi(\ell(u) - \beta_0)], & p \geq p^* \end{cases}$$

where

$$p^* \equiv \Phi(\ell(u) - \beta_0) / P_D(A)$$

$$P_D(A) \equiv 1 - \Phi(r(u) - \beta_0) + \Phi(\ell(u) - \beta_0)$$

$$\ell(u) \equiv (-c - u) / \rho$$

$$r(u) \equiv (c - u) / \rho$$

## The Distribution of $Y_\beta | (A^c, U = u)$

$$\begin{bmatrix} Y_\beta \\ Y_\delta \end{bmatrix} | (U = u) = \begin{bmatrix} \beta + Z \\ u + \rho(\beta + Z) \end{bmatrix}, \quad Z \sim N(0, 1)$$

If  $A$  does not occur, when we select  $M2$ . Under  $H_0: \beta = \beta_0$ ,

$$\begin{aligned} \mathbb{P}_{\beta_0}(Y_\beta \leq y | A^c, U = u) &= \frac{\mathbb{P}_{\beta_0}(\{Y_\beta \leq y\} \cap A^c | U = u)}{\mathbb{P}_{\beta_0}(A^c | U = u)} \\ &= \frac{\mathbb{P}(\{Z \leq y - \beta_0\} \cap \{|u + \rho(\beta_0 + Z)| < c\})}{\mathbb{P}(|u + \rho(\beta_0 + Z)| < c)} \end{aligned}$$

$$F_{\beta_0}(y|A^c, U = u)$$

As above, define  $\ell(u) = (-c - u)/\rho$ ,  $r(u) = (c - u)/\rho$ . We have:

$$F_{\beta_0}(y|A^c, U = u) = P_N(A^c)/P_D(A^c)$$

where

$$P_D(A^c) \equiv \Phi(r(u) - \beta_0) - \Phi(\ell(u) - \beta_0)$$

$$P_N(A^c) \equiv \begin{cases} 0, & y < \ell(u) \\ \Phi(y - \beta_0) - \Phi(\ell(u) - \beta_0), & \ell(u) \leq y \leq r(u) \\ \Phi(r(u) - \beta_0) - \Phi(\ell(u) - \beta_0), & y > r(u) \end{cases}$$

Notice that this is a CDF with a bounded support set:  $y \in [\ell(u), r(u)]$



$$Q_{\beta_0}(p|A^c, U = u)$$

Inverting the CDF from the preceding slide:

$$Q_{\beta_0}(p|A^c, U = u) = \beta_0 + \Phi^{-1} [p \times P_D(A^c) + \Phi(\ell(u) - \beta_0)]$$

where:

$$P_D(A^c) \equiv \Phi(r(u) - \beta_0) - \Phi(\ell(u) - \beta_0)$$

$$\ell(u) \equiv (-c - u)/\rho$$

$$r(u) \equiv (c - u)/\rho$$

# Equal-tailed Selective Test

## Conditional on $A$

1. Compute observed value  $u$  of  $U = Y_\delta - \rho Y_\beta$  (given  $A$ ).
2. Compute  $q_{\alpha/2} \equiv Q_{\beta_0}(\alpha/2|A, U = u)$
3.  $q_{1-\alpha/2} \equiv Q_{\beta_0}(1 - \alpha/2|A, U = u)$
4. Reject  $H_0: \beta = \beta_0$  if  $Y_\beta$  lies outside outside  $[q_{\alpha/2}, q_{1-\alpha/2}]$ .

## Conditional on $A^c$

Same as above, but replace  $A$  with  $A^c$  in the preceding expressions.

## Constructing a Confidence Interval

Simply invert the test: find the values of  $\beta_0$  that are not rejected.

*Valid conditional on  $(U = u) \implies$  valid unconditionally!*