

# Econ 722 – Advanced Econometrics IV, Part II

Francis J. DiTraglia

University of Pennsylvania

# Lecture #1 – AIC-type Information Criteria

Kullback-Leibler Divergence

Bias of Maximized Sample Log-Likelihood

Review of Asymptotics for Mis-specified MLE

Deriving AIC and TIC

Corrected AIC ( $AIC_c$ )

# Kullback-Leibler (KL) Divergence

## Motivation

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

## Definition

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

## Properties

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

# KL Divergence and Mis-specified MLE

Pseudo-true Parameter Value  $\theta_0$

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

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

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

Goal: Compare Mis-specified Models

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

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

# How to Estimate Expected Log Likelihood?

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

## Unbiased but Infeasible

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

## Biased but Feasible

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

## Intuition for the Bias

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

# What to do about this bias?

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

## Tradeoffs

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

# Recall: Asymptotics for Mis-specified ML Estimation

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

## Fundamental Expansion

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

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

## Central Limit Theorem

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

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

## Information Matrix Equality

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

# Bias Relative to Infeasible Plug-in Estimator

## Definition of Bias Term $B$

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

## 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

# Lecture #2 – More on “Classical” Model Selection

Mallow's  $C_p$

Bayesian Model Comparison

Laplace Approximation

Bayesian Information Criterion (BIC)

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

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

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

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



# Operationalizing the Bias-Variance Tradeoff Idea

## Mallow's $C_p$

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

## Notation

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

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

Why not compare  $RSS(m)$ ?

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

From your Problem Set

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

Bias-Variance Tradeoff

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

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

Step 1: Algebra

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

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

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

# 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  $\mathbf{X}$

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

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

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

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

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

### Bias-Variance Tradeoff

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

### Mallow's $C_p$

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

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

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

## Notation

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

## Crucial Fact

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

## Step 4: Algebra using the preceding fact

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

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

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

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

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

### Bias of Plug-in Estimator

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

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

Want An Unbiased Estimator of This:

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

Previous Slide:

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

End Result:

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

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



## Why is this different from the textbook formula?

Just algebra, but tedious. . .

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

Therefore:

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

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

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

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

# Bayesian Model Comparison: Marginal Likelihoods

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

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

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

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

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

# Laplace (aka Saddlepoint) Approximation

Suppress model index  $m$  for simplicity.

General Case: for  $T$  large...

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

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

Use to Approximate Marginal Likelihood

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

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

# Laplace Approximation to Marginal Likelihood

Suppress model index  $m$  for simplicity.

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

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

# Bayesian Information Criterion

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

Take Logs and Multiply by 2

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

The BIC

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

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

large-sample Frequentist approx. to Bayesian marginal likelihood

# Lecture #3 – Cross-Validation

Model selection via a Hold-out Sample

K-fold Cross-validation

Asymptotic Equivalence Between LOO-CV and TIC

Influence Functions

# 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  $\hat{\theta}_{(t)}$  around  $\hat{\theta}$ :

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

Crucial point: the first-order term is not zero in this case. (Why?)

## Overview of Proof

From expansion on previous slide, 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

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

# Digression: Functionals and Influence Functions

## (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  $\mathbb{T}$ ?

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

## Back to the Proof...

### Step 1

The influence function for ML estimation turns out to be

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

### Step 2

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

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

### Step 3

Evaluating Step 1 at  $\hat{G}$  and substituting into Step 2

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

# Lecture #4 – Asymptotic Properties

Overview

Weak Consistency

Consistency

Efficiency

AIC versus BIC in a Simple Example

# Overview

- ▶ What happens as  $T \rightarrow \infty$ ?
- ▶ Consistency: choose “best” model wpa 1
- ▶ Efficiency: procedure with good risk properties
- ▶ Can't have both at once.
- ▶ Large, fairly technical literature: only a brief overview today.
- ▶ More details: 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

- ▶ Roughly speaking, a model selection criterion is called efficient if it performs “nearly as well” as the theoretical optimum relative to some loss function.
- ▶ More broadly, an efficient/conservative criterion is one that has “good risk properties.”
- ▶ We don’t have time to go into detail, so we’ll look at a particular example. . .

# Consistency versus 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

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

$$\begin{aligned}\ell_T(\mu) &= \sum_{t=1}^T \log \left( \frac{1}{2\pi} \exp \left\{ -\frac{1}{2} (Y_t - \mu)^2 \right\} \right) \\ &= -\frac{T}{2} \log(2\pi) - \frac{1}{2} \sum_{t=1}^T (Y_t - \mu)^2 \\ &= -\frac{T}{2} \log(2\pi) - \frac{T}{2} \hat{\sigma}^2 - \frac{T}{2} (\bar{Y} - \mu)^2 \\ &= \text{Constant} - \frac{T}{2} (\bar{Y} - \mu)^2\end{aligned}$$

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

$$\begin{aligned} T\hat{\sigma}^2 &= \sum_{t=1}^T (Y_t - \bar{Y})^2 = \sum_{t=1}^T (Y_t - \mu + \mu - \bar{Y})^2 = \sum_{t=1}^T [(Y_t - \mu) - (\bar{Y} - \mu)]^2 \\ &= \sum_{t=1}^T (Y_t - \mu)^2 - \sum_{t=1}^T 2(Y_t - \mu)(\bar{Y} - \mu) + \sum_{t=1}^T (\bar{Y} - \mu)^2 \\ &= \left[ \sum_{t=1}^T (Y_t - \mu)^2 \right] - 2(\bar{Y} - \mu) \left( \sum_{t=1}^T Y_t - \sum_{t=1}^T \mu \right) + T(\bar{Y} - \mu)^2 \\ &= \left[ \sum_{t=1}^T (Y_t - \mu)^2 \right] - 2(\bar{Y} - \mu)(T\bar{Y} - T\mu) + T(\bar{Y} - \mu)^2 \\ &= \left[ \sum_{t=1}^T (Y_t - \mu)^2 \right] - 2T(\bar{Y} - \mu)^2 + T(\bar{Y} - \mu)^2 \\ &= \left[ \sum_{t=1}^T (Y_t - \mu)^2 \right] - T(\bar{Y} - \mu)^2 \end{aligned}$$

# 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

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



## Case I: $\mu \neq 0$

Apply theory from earlier in lecture...

### KL-Divergence of $M_1$

$M_1$  is the true DGP with minimized KL-divergence equal to zero.

### KL-Divergence of $M_0$

- ▶ Truth:  $g(y) = (2\pi)^{-1/2} \exp \{-(y - \mu)^2/2\}$
- ▶  $M_0$ :  $f(y) = (2\pi)^{-1/2} \exp \{-y^2/2\}$
- ▶ Hence:  $\log g(y) - \log f(y) = -\frac{1}{2}(y - \mu)^2 + \frac{1}{2}y^2 = \mu(y - \frac{\mu}{2})$

$$\begin{aligned} \text{KL}(g; M_0) &= \int_{\mathbb{R}} \mu(y - \mu/2)(2\pi)^{-1/2} \exp \{(y - \mu)^2/2\} dy \\ &= \mu(\mu - \mu/2) = \mu^2/2 \end{aligned}$$

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

### Condition on KL-Divergence

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

### Condition on Penalty

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

## Case II: $\mu = 0$

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  $\sqrt{2}$ :

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

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  $T$  since Var. of well-behaved estimator  $= O(1/T)$

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

# Simplifying the MSE Risk Function

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

Let  $X = \mathbf{1}\{A\}$  where  $A = \left\{|\sqrt{T}\mu + Z| \geq \sqrt{dT}\right\}$

$$\begin{aligned}R_T(\mu) &= \mathbb{E} \left[ \left( \sqrt{T}\hat{\mu} - \sqrt{T}\mu \right)^2 \right] \\&= \mathbb{E} \left\{ \left[ \left( \sqrt{T}\mu + Z \right) X - \sqrt{T}\mu \right]^2 \right\} \\&= \mathbb{P}(A) \mathbb{E} \left\{ \left[ \left( \sqrt{T}\mu + Z \right) - \sqrt{T}\mu \right]^2 \middle| X = 1 \right\} + [1 - \mathbb{P}(A)] \left( \sqrt{T}\mu \right)^2 \\&= \mathbb{P}(A) \mathbb{E} \left[ Z^2 | X = 1 \right] + [1 - \mathbb{P}(A)] T\mu^2\end{aligned}$$

So we need to calculate  $\mathbb{P}(A) \mathbb{E}[Z^2 | X = 1]$  and  $\mathbb{P}(A)$ .



## Calculating $\mathbb{P}(A)$

Define  $a = (-\sqrt{d_T} - \sqrt{T}\mu)$  and  $b = (\sqrt{d_T} - \sqrt{T}\mu)$

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}\left(|\sqrt{T}\mu + Z| \geq \sqrt{d_T}\right) \\&= \mathbb{P}\left(\sqrt{T}\mu + Z \geq \sqrt{d_T}\right) + \mathbb{P}\left(\sqrt{T}\mu + Z \leq -\sqrt{d_T}\right) \\&= \mathbb{P}(Z \geq b) + \mathbb{P}(Z \leq a) \\&= 1 - \Phi(b) + \Phi(a)\end{aligned}$$

And hence:

$$1 - \mathbb{P}(A) = \Phi(b) - \Phi(a)$$

## Calculating $\mathbb{P}(A) \mathbb{E}[Z^2|X = 1]$ – Step 1

### Conditional Density of $Z|X = 1$

$$f(z|x = 1) = \frac{\mathbf{1}(A)\varphi(z)}{\mathbb{P}(A)} \quad \text{where } \varphi \text{ is the } N(0, 1) \text{ density}$$

Therefore:

$$\begin{aligned}\mathbb{P}(A) \mathbb{E}[Z^2|X = 1] &= \mathbb{P}(A) \int_{\mathbb{R}} z^2 \left[ \frac{\mathbf{1}(A)\varphi(z)}{\mathbb{P}(A)} \right] dz \\ &= \int_{-\infty}^a z^2 \varphi(z) dz + \int_b^{\infty} z^2 \varphi(z) dz\end{aligned}$$

## Calculating $\mathbb{P}(A) \mathbb{E}[Z^2|X = 1]$ – Step 2

Unconditional Expectation:  $\mathbb{E}[Z^2]$

$$1 = \mathbb{E}[Z^2] = \int_{-\infty}^a z^2 \varphi(z) \, dz + \int_a^b z^2 \varphi(z) \, dz + \int_b^{\infty} z^2 \varphi(z) \, dz$$

Therefore:

$$\begin{aligned} \mathbb{P}(A) \mathbb{E}[Z^2|X = 1] &= \int_{-\infty}^a z^2 \varphi(z) \, dz + \int_b^{\infty} z^2 \varphi(z) \, dz \\ &= 1 - \int_a^b z^2 \varphi(z) \, dz \end{aligned}$$

## Calculating $\mathbb{P}(A) \mathbb{E}[Z^2|X = 1]$ – Step 3

### Integration By Parts

Take  $u = -z$  and  $dv = -z \exp\{-z^2/2\}$  since

$$\frac{d}{dz} (\exp\{-z^2/2\}) = -z \exp\{-z^2/2\}$$

Thus,  $v = \exp\{-z^2/2\}$ ,  $du = -1$  and

$$\begin{aligned} \int_a^b z^2 \phi(z) dz &= (2\pi)^{-1/2} \int_a^b z^2 \exp\{-z^2/2\} dz \\ &= (2\pi)^{-1/2} \left[ -z \exp\{-z^2/2\} \Big|_a^b + \int_a^b \exp\left\{-\frac{z^2}{2}\right\} dz \right] \\ &= a\phi(a) - b\phi(b) + \Phi(b) - \Phi(a) \end{aligned}$$

# The Simplified MSE Risk Function

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

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/)

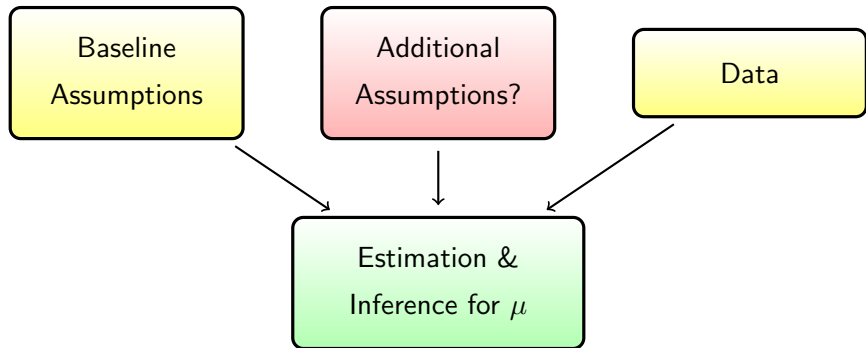
## Punchline: Risk of the Post-Selection Estimator

- ▶ AIC: bounded worst-case risk
- ▶ BIC: low risk in a neighborhood of  $\mu = 0$  in exchange for **unbounded** worst-case risk as sample size grows
- ▶ General phenomenon: consistency and efficiency are mutually exclusive: consistent criteria have unbounded worst-case risk.
- ▶ For more details, see Yang (2007, ET)

# Lecture #7 – Focused Moment Selection

DiTraglia (2016, JoE)

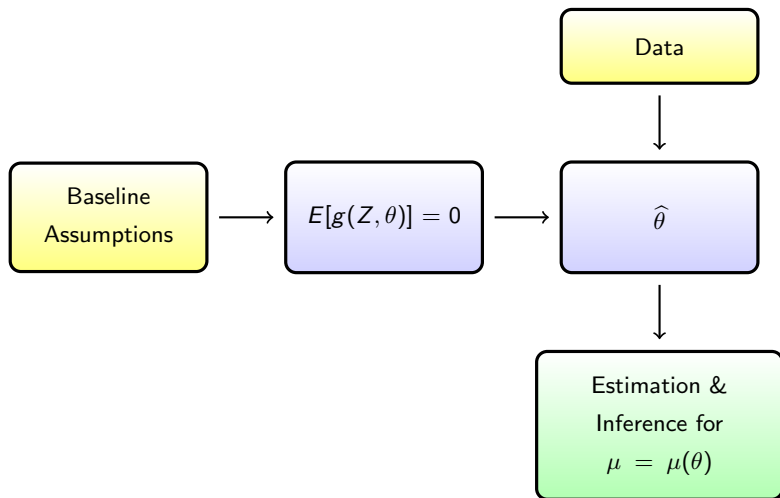
# Focused Moment Selection Criterion (FMSC)



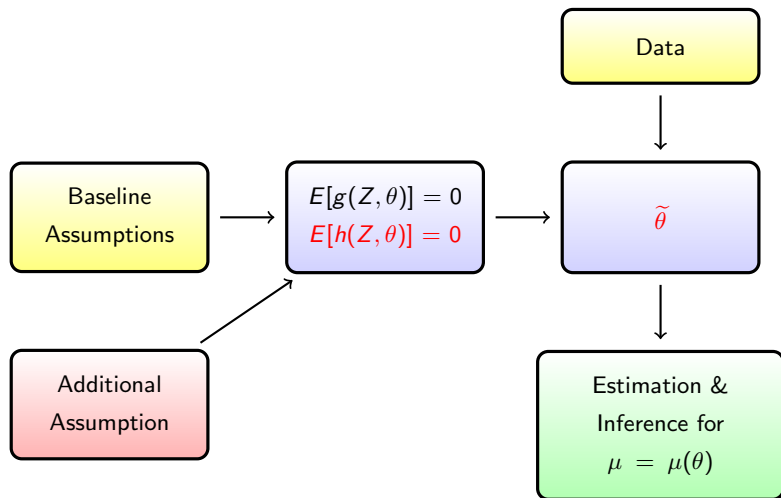
1. Choose False Assumptions on Purpose
2. Focused Choice of Assumptions
3. Local mis-specification
4. Averaging, Inference post-selection



# 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

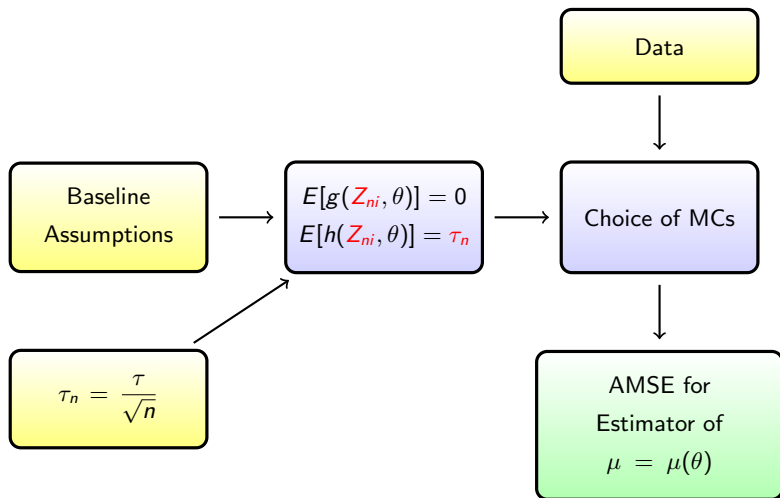
$$y_i = \beta x_i + \epsilon_i$$

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

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

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

# 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

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

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

$$E[\mathbf{z}_i^{(1)} v_i] = \tau / \sqrt{n}$$

# 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 $\mu$

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

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

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

## Delta Method

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

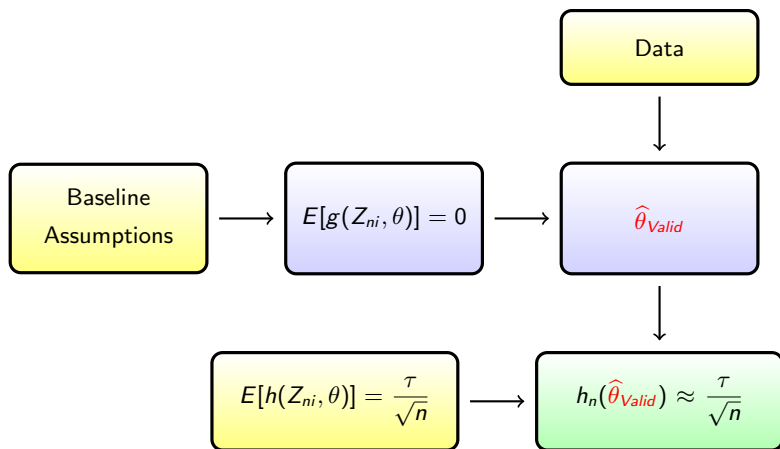
FMSC: Estimate  $\text{AMSE}(\hat{\mu}_S)$  and minimize over  $S$

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

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

$$y_i = 0.5x_i + \epsilon_i$$

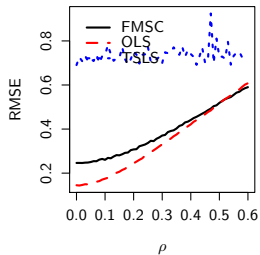
$$x_i = \pi(z_{1i} + z_{2i} + z_{3i}) + v_i$$

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

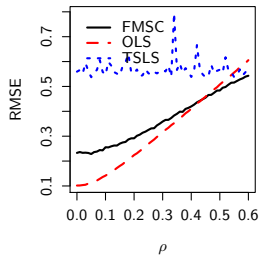
$$\mathcal{S} = \begin{bmatrix} 1 & \rho & 0 & 0 & 0 \\ \rho & 1 - \pi^2 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/3 \end{bmatrix}$$

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

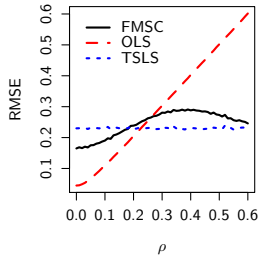
$N = 50, \pi = 0.2$



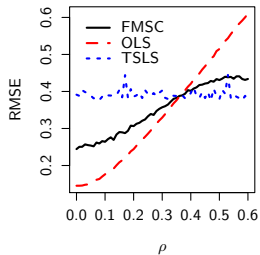
$N = 100, \pi = 0.2$



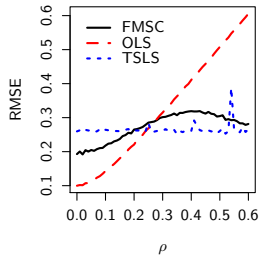
$N = 500, \pi = 0.2$



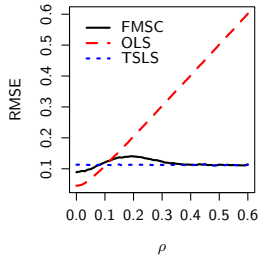
$N = 50, \pi = 0.4$



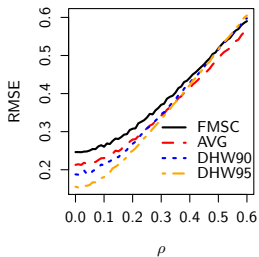
$N = 100, \pi = 0.4$



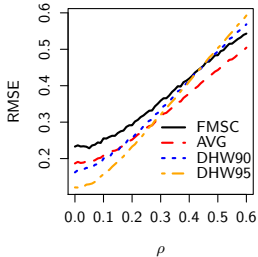
$N = 500, \pi = 0.4$



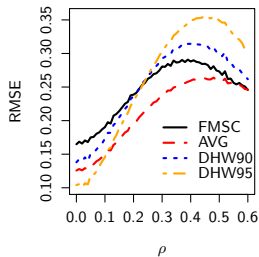
$N = 50, \pi = 0.2$



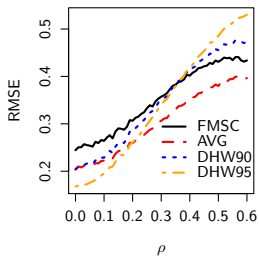
$N = 100, \pi = 0.2$



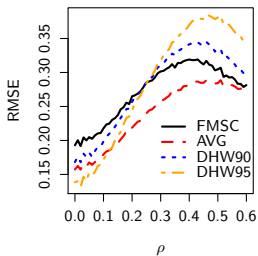
$N = 500, \pi = 0.2$



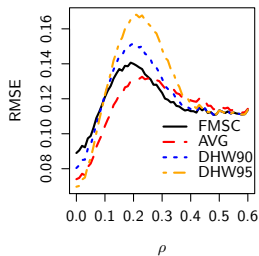
$N = 50, \pi = 0.4$



$N = 100, \pi = 0.4$



$N = 500, \pi = 0.4$



## Choosing Instrumental Variables Simulation

$$y_i = 0.5x_i + \epsilon_i$$

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

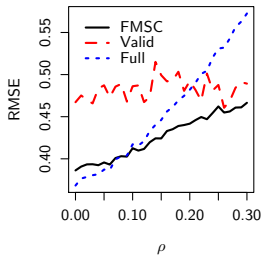
$$(\epsilon_i, v_i, w_i, z_{1i}, z_{2i}, z_{3i})' \sim \text{iid } N(0, \mathcal{V})$$

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

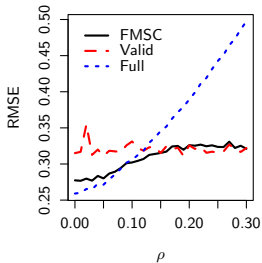
$$\gamma = \text{Cor}(x, w), \quad \rho = \text{Cor}(w, \epsilon), \quad \text{First-Stage } R^2 = 1/9 + \gamma^2$$

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

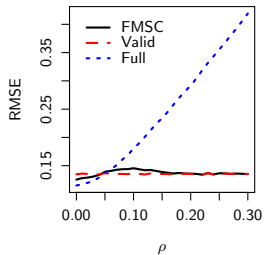
$N = 50, \gamma = 0.2$



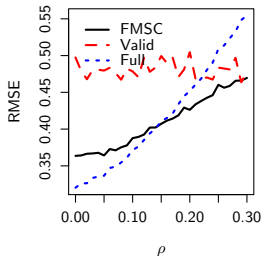
$N = 100, \gamma = 0.2$



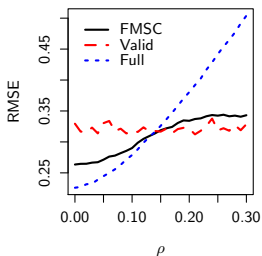
$N = 500, \gamma = 0.2$



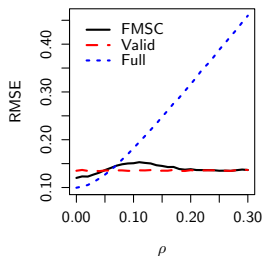
$N = 50, \gamma = 0.3$



$N = 100, \gamma = 0.3$



$N = 500, \gamma = 0.3$



# Alternative Moment Selection Procedures

## Downward $J$ -test

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

## Andrews (1999) – GMM Moment Selection Criteria

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

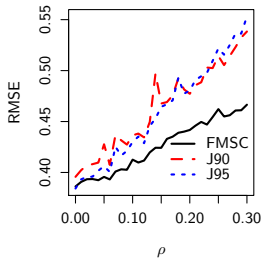
## Hall & Peixe (2003) – Canonical Correlations Info. Criterion

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

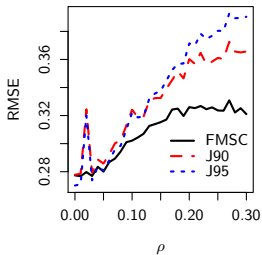
## Penalty/Bonus Terms

Analogies to AIC, BIC, and Hannan-Quinn

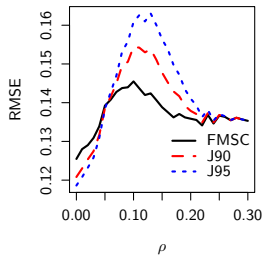
$N = 50, \gamma = 0.2$



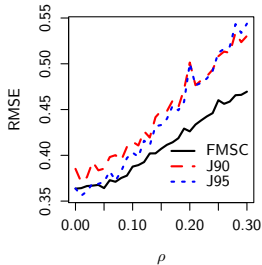
$N = 100, \gamma = 0.2$



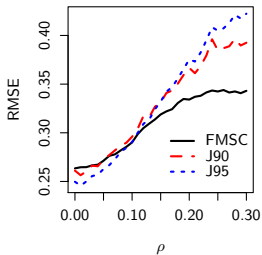
$N = 500, \gamma = 0.2$



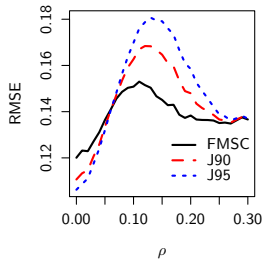
$N = 50, \gamma = 0.3$



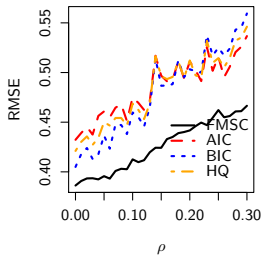
$N = 100, \gamma = 0.3$



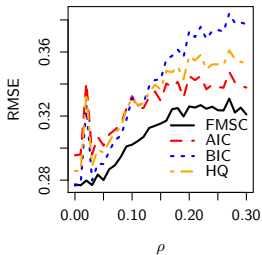
$N = 500, \gamma = 0.3$



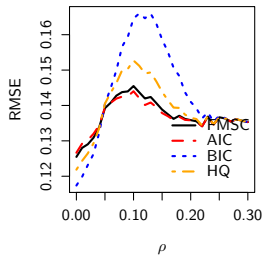
$N = 50, \gamma = 0.2$



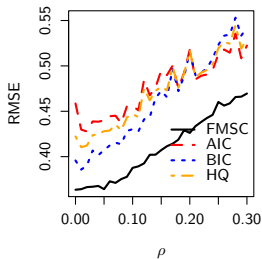
$N = 100, \gamma = 0.2$



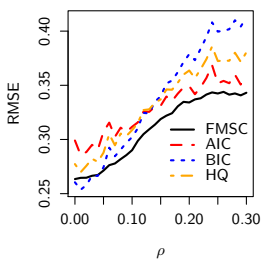
$N = 500, \gamma = 0.2$



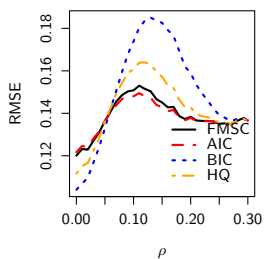
$N = 50, \gamma = 0.3$



$N = 100, \gamma = 0.3$



$N = 500, \gamma = 0.3$





# 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 #8 – High-Dimensional Regression I

The James-Stein Estimator

QR Decomposition

Singular Value Decomposition

Review of Principal Component Analysis (PCA)

Ridge Regression

# 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?

# Dominance and Admissibility

## Notation

Let  $R$  be a risk function, e.g. MSE, and  $\hat{\theta}$  and  $\tilde{\theta}$  be estimators of  $\theta$ .

## Dominance

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

## Admissibility

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

## Inadmissibility

To prove that an estimator  $\tilde{\theta}$  is **inadmissible** it suffices to find an estimator  $\hat{\theta}$  that dominates it.



## A Very Simple Example: $X \sim N(\theta, I)$

### Goal

Estimate the  $p$ -vector of unknown parameters  $\theta$  using  $X$ .

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

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

### MSE 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  $MSE(\hat{\theta}) = p$ .

## A Very Simple Example: $X \sim N(\theta, I)$

### 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

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

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

# 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 $p \geq 3$

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

- ▶  $\mathbb{E}[1/(X'X)]$  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!

# 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

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

- ▶  $U = m \times m$  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 positive eigenvalues of  $A'A$  which are *identical* to the positive eigenvalues of  $AA'$ .
- ▶ (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  $D = \Lambda^2$  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  $i$

# Getting the PC Scores for Free from the SVD

PC scores for  $j$ th PC

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

# Ridge Regression