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

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

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 θ_0

$$\widehat{\theta}_{MLE} \xrightarrow{p} \theta_0 \equiv \underset{\theta \in \Theta}{\arg\min}\ \mathrm{KL}(g; f_\theta) = \underset{\theta \in \Theta}{\arg\max}\ \mathbb{E}_G[\log f(Y|\theta)]$$

What if f_θ is correctly specified?

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

Goal: Compare Mis-specified Models

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

where θ_0 is the pseudo-true parameter value for f_θ and γ_0 is the pseudo-true parameter value for h_γ.

How to Estimate Expected Log Likelihood?

For simplicity: $Y_1, \ldots, Y_n \sim$ iid $g(y)$

Unbiased but Infeasible

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

Biased but Feasible

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

Intuition for the Bias

$T^{-1}\ell(\widehat{\theta}_{MLE}) > T^{-1}\ell(\theta_0)$ unless $\widehat{\theta}_{MLE} = \theta_0$. 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 θ_0. For simplicity $Y_1, \ldots, Y_T \sim$ iid $g(y)$.

Fundamental Expansion

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

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

Central Limit Theorem

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

$$\sqrt{T}(\widehat{\theta} - \theta_0) \to_d J^{-1}U \sim N_p(0, J^{-1}KJ^{-1})$$

Information Matrix Equality

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

Bias Relative to Infeasible Plug-in Estimator

Definition of Bias Term B

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

Question to Answer

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

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

Derivation of AIC/TIC

Step 1: Taylor Expansion

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

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

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

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

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

$$T(\widehat{\theta} - \theta_0)' J(\widehat{\theta} - \theta_0) \to_d U' J^{-1} U$$

Derivation of AIC/TIC Continued...

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

$$T(\widehat{\theta} - \theta_0)' J(\widehat{\theta} - \theta_0) \to_d U' J^{-1} U$$

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

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

Final Result:

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

TIC and AIC

Takeuchi's Information Criterion

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

Akaike's Information Criterion

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

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, $\ell(\widehat{\theta})$ 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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma_0^2 \mathbf{I}_T), \quad k \text{ Regressors}$$

Can Show That

$$KL(g, f) = \frac{T}{2}\left[\frac{\sigma_0^2}{\sigma_1^2} - \log\left(\frac{\sigma_0^2}{\sigma_1^2}\right) - 1\right] + \left(\frac{1}{2\sigma_1^2}\right)(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_1)'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_1)$$

Where f is a normal regression model with parameters $(\boldsymbol{\beta}_1, \sigma_1^2)$ that might not be the true parameters.

But how can we use this?

$$KL(g, f) = \frac{T}{2}\left[\frac{\sigma_0^2}{\sigma_1^2} - \log\left(\frac{\sigma_0^2}{\sigma_1^2}\right) - 1\right] + \left(\frac{1}{2\sigma_1^2}\right)(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_1)'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_1)$$

1. Would need to know (β_1, σ_1^2) for candidate model.
   - Easy: just use MLE $(\widehat{\boldsymbol{\beta}}_1, \widehat{\sigma}_1^2)$
2. Would need to know $(\boldsymbol{\beta}_0, \sigma_0^2)$ 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 $(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2)$ is consistent for truth.

Deriving the Corrected AIC

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

$$\widehat{KL} = \frac{T}{2}\left[\frac{\sigma_0^2}{\widehat{\sigma}^2} - \log\left(\frac{\sigma_0^2}{\widehat{\sigma}^2}\right) - 1\right] + \left(\frac{1}{2\widehat{\sigma}^2}\right)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$$

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

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

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

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

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

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

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

- 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 $\boldsymbol{\beta}$ is known.

Notation

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

In-sample versus Out-of-sample Prediction Error

Why not compare RSS(m)?

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

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 $\mathbf{X}\widehat{\boldsymbol{\beta}}_m$ relative to infeasible optimum $\boldsymbol{X}\beta$

Step 1: Algebra

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

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

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

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

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

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

where k_m denotes the number of regressors in $\mathbf{X}_m$ and
$\text{tr}(\mathbf{P}_m) = \text{tr}\left\{\mathbf{X}_m\left(\mathbf{X}_m'\mathbf{X}_m\right)^{-1}\mathbf{X}_m'\right\} = \text{tr}\left\{\mathbf{X}_m'\mathbf{X}_m\left(\mathbf{X}_m'\mathbf{X}_m\right)^{-1}\right\} = \text{tr}(\mathbf{I}_m)$

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

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

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 $\boldsymbol{\beta}$ and σ^2.
- Solution: Mallow's C_p constructs an unbiased estimator.
- Idea: what about plugging in $\widehat{\boldsymbol{\beta}}$ to estimate second term?

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

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

Notation

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

Crucial Fact

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

Step 4: Algebra using the preceding fact

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

Substituting $\widehat{\beta}$ doesn't work...

Step 5: Use "Trace Trick" on second term from Step 4

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

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

Bias of Plug-in Estimator

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

Putting Everything Together: Mallow'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:

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

End Result:

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

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

Why is this different from the textbook formula?

Just algebra, but tedious...

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

Therefore:

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

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

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

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

Bayesian Model Comparison: Marginal Likelihoods

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

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

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

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

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

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...

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

$$p = \dim(\boldsymbol{\theta}), \;\; \boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta}\in\Theta} h(\boldsymbol{\theta}), \;\; H(\theta_0) = -\left.\frac{\partial^2 h(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$$

Use to Approximate Marginal Likelihood

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

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

Laplace Approximation to Marginal Likelihood

Suppress model index m for simplicity.

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

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

Bayesian Information Criterion

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

Take Logs and Multiply by 2

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

The BIC

Assume uniform prior over models and ignore lower order terms:

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

large-sample Frequentist approx. to Bayesian marginal likelihood

Lecture #3 – Cross-Validation

Model selection via a Hold-out Sample

K-fold Cross-validation

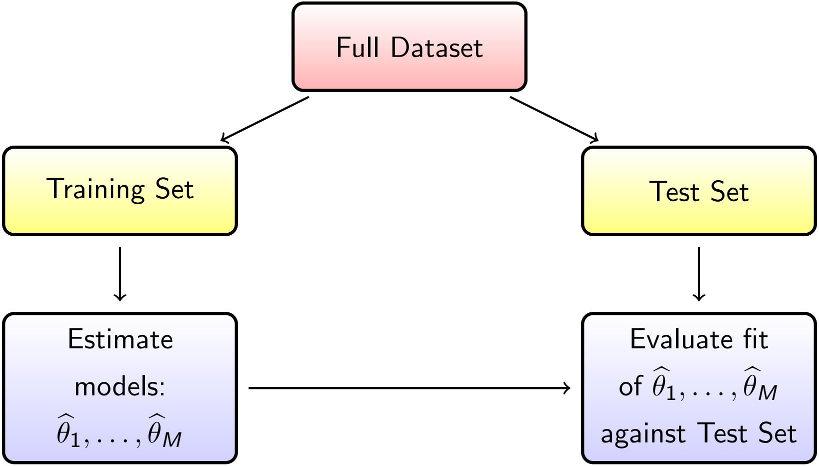
Asymptotic Equivalence Between LOO-CV and TIC

Influence Functions

Model Selection using a Hold-out Sample

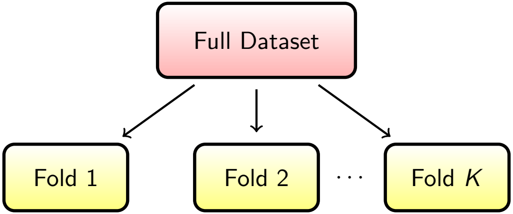
- The real problem is double use of the data: first for estimation, then for model comparison.
  - Maximized sample log-likelihood is an overly optimistic estimate of expected log-likelihood and hence KL-divergence
  - In-sample squared prediction error is an overly optimistic estimator of out-of-sample 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^{th} fold as a hold-out sample and estimate model using all observations except those in fold k: yielding estimator $\widehat{\theta}(-k)$.

K-fold Cross-Validation: "Pseudo-out-of-sample"

Step 2

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

Step 3

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

Step 4

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

K-fold Cross-Validation: "Pseudo-out-of-sample"

Step 4

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

Step 5

Define $\text{CV}_K = \frac{1}{T}\sum_{t=1}^{T} L\left(y_t, \widehat{y}_t^{-k(t)}\right)$ where L is a loss function.

Step 5

Repeat for each model & choose m to minimize $\text{CV}_K(m)$.

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 $Y_1, \ldots, Y_T \sim$ iid. Let $\widehat{\theta}_{(t)}$ be the maximum likelihood estimator based on all observations except t and $\widehat{\theta}$ be the full-sample estimator.

Log-likelihood as "Loss"

$\text{CV}_1 = \frac{1}{T}\sum_{t=1}^{T} \log f(y_t|\widehat{\theta}_{(t)})$ but since min. KL = max. log-like. we choose the model with highest $\text{CV}_1(m)$.

Overview of the Proof

First-Order Taylor Expansion of $\widehat{\theta}_{(t)}$ around $\widehat{\theta}$:

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

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:

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

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

Overview of Proof

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

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

So it suffices to show that

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

Digression: Functionals and Influence Functions

(Statistical) Functional

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

Example: ML Estimation

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

Influence Function

Let δ_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 $\mathbb{T}$?

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

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 $\widehat{G}$ denote the empirical CDF based on $y_1, \ldots, y_T$. Then:

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

Step 3

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

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

Lecture #4 – Asymptotic Properties

Overview

Weak Consistency

Consistency

Efficiency

AIC versus BIC in a Simple Example

Overview

- What happens as $T \to \infty$?
- 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:

$$
\begin{aligned}
TIC &= 2\ell_T(\widehat{\theta}) - \text{trace}\left\{\widehat{J}^{-1}\widehat{K}\right\} \\
AIC &= 2\ell_T(\widehat{\theta}) - 2\ \text{length}(\theta) \\
BIC &= 2\ell_T(\widehat{\theta}) - \log(T)\ \text{length}(\theta)
\end{aligned}
$$

Generic penalty $c_{T,k}$

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

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

Weak Consistency: Suppose M_{k_0} Uniquely Minimizes KL

Assumption

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

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 $T^{-1} c_{T,k} \xrightarrow{p} 0$.

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

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

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

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

Sufficient Conditions

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

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

where θ_k^* and θ_ℓ^* are the KL minimizing parameter values.

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

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

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:

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

- Term in braces is *positive* since $\ell \in \mathcal{J} \backslash \mathcal{J}_0$, i.e. ℓ is not as parsimonious as j_0
- $\log(T) \to \infty$, 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 $\mathrm{IC}_m = 2\ell(\theta) - d_T \times \mathrm{length}(\theta)$.

True DGP

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

Candidate Models

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

$$\mathrm{IC}_0 = 2 \max_{\mu} \{\ell(\mu) \colon \mathrm{M}_0\}$$

$$\mathrm{IC}_1 = 2 \max_{\mu} \{\ell(\mu) \colon \mathrm{M}_1\} - d_T$$

Log-Likelihood Function

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

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

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

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

The Selected Model $\widehat{\mathsf{M}}$

Information Criteria

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

$$\mathsf{IC}_0 = 2 \max_{\mu} \{\ell(\mu) \colon \mathsf{M}_0\} = 2 \times \text{Constant} - T\bar{Y}^2$$

$$\mathsf{IC}_1 = 2 \max_{\mu} \{\ell(\mu) \colon \mathsf{M}_1\} - d_T = 2 \times \text{Constant} - d_T$$

Difference of Criteria

$$\mathsf{IC}_1 - \mathsf{IC}_0 = T\bar{Y}^2 - d_T$$

Selected Model

$$\widehat{\mathsf{M}} = \begin{cases} \mathsf{M}_1, & |\sqrt{T}\bar{Y}| \geq \sqrt{d_T} \\ \mathsf{M}_0, & |\sqrt{T}\bar{Y}| < \sqrt{d_T} \end{cases}$$

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

$$
\begin{aligned}
\mathrm{KL}(g; \mathrm{M}_0) &= \int_{\mathbb{R}} \mu(y - \mu/2)(2\pi)^{-1/2} \exp\left\{(y-\mu)^2/2\right\} \, \mathrm{d}y \\
&= \mu(\mu - \mu/2) = \mu^2/2
\end{aligned}
$$

Verifying Weak Consistency: $\mu \neq 0$

Condition on KL-Divergence

$$\liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \{KL(g;\mathrm{M}_0) - KL(g;\mathrm{M}_1)\} = \liminf_{n\to\infty} \frac{1}{T} \sum_{t=1}^{T} \left(\frac{\mu^2}{2} - 0\right) > 0$$

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 $T \to \infty$.

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 $\mathbb{P}(c_{T,k} - c_{T,0}) \to \infty$
- BIC satisfies this but AIC doesn't.

Finite-Sample Selection Probabilities: AIC

AIC Sets $d_T = 2$

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

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

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

Finite-Sample Selection Probabilities: BIC

BIC sets $d_T = \log(T)$

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

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

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

Interactive Demo: AIC vs BIC

`https://fditraglia.shinyapps.io/CH_Figure_4_1/`

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):

$$\begin{aligned} P\left(\widehat{M} = M_1\right) &= P\left(|\sqrt{T}\bar{Y}| \geq \sqrt{d_T}\right) = P(|Z| \geq \sqrt{d_T}) \\ &= P(Z^2 \geq d_T) = P(\chi^2_1 \geq d_T) \end{aligned}$$

- AIC: $d_T = 2$ and $P(\chi^2_1 \geq 2) \approx 0.157$.
- BIC: $d_T = \log(T)$ and $P(\chi^2_1 \geq \log T) \rightarrow 0$ 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

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

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 $= O(1/T)$

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

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 $A = \left\{|\sqrt{T}\mu + Z| \geq \sqrt{d_T}\right\}$

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

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:

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

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

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

$$1 = \mathbb{E}[Z^2] = \int_{-\infty}^{a} z^2\varphi(z)\ \mathrm{d}z + \int_{a}^{b} z^2\varphi(z)\ \mathrm{d}z + \int_{b}^{\infty} z^2\varphi(z)\ \mathrm{d}z$$

Therefore:

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

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

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

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

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

The Simplifed 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

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

`https://fditraglia.shinyapps.io/CH_Figure_4_2/`

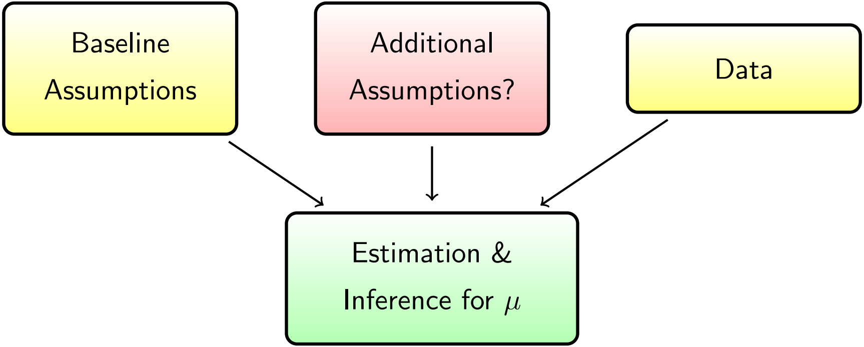
Punchline: Risk of the Post-Selection Estimator

- AIC: bounded worst-case risk
- BIC: low risk in a neighborhood of $\mu = 0$ in exhange 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

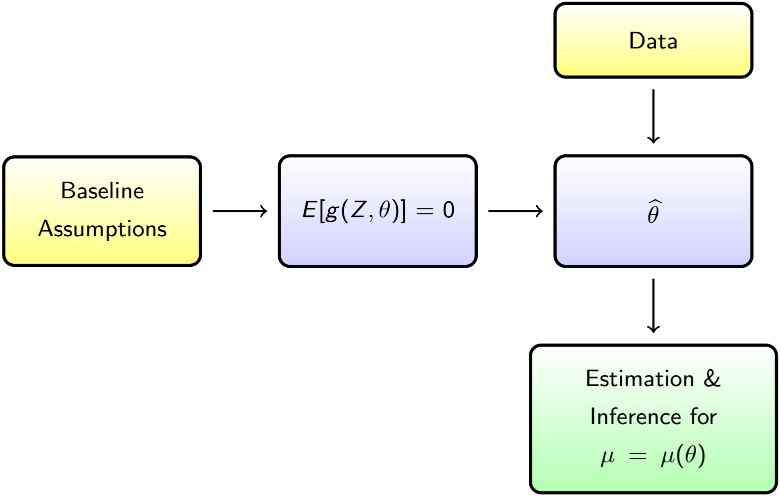
DiTraglia (2016, JoE)

Focused Moment Selection Criterion (FMSC)

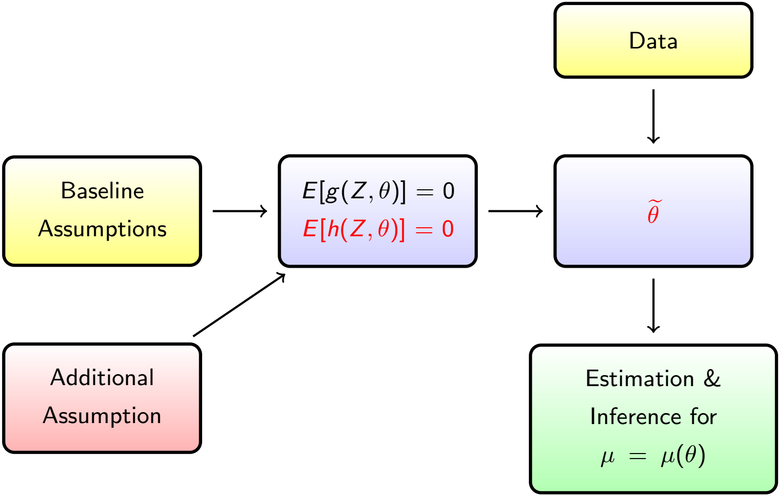


1. Choose False Assumptions on Purpose
2. Focused Choice of Assumptions
3. Local mis-specification
4. Averaging, Inference post-selection

GMM Framework



Adding Moment Conditions



Ordinary versus Two-Stage Least Squares

$$
\begin{aligned}
y_i &= \beta x_i + \epsilon_i \\
x_i &= \mathbf{z}_i' \boldsymbol{\pi} + v_i
\end{aligned}
$$

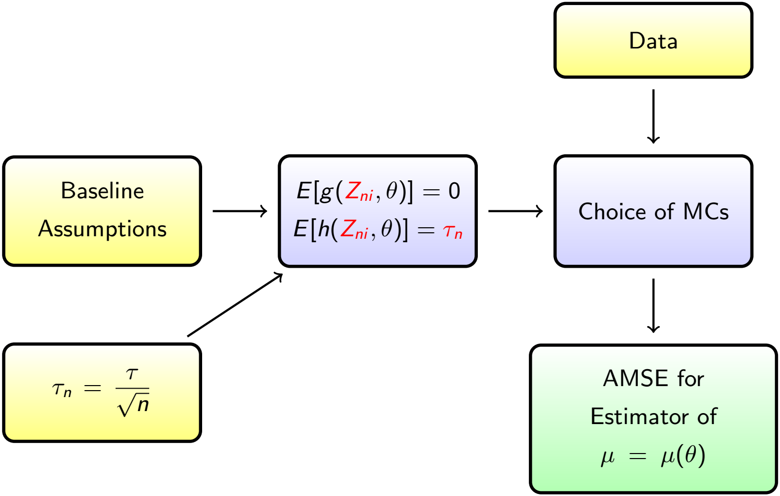
$$
\begin{aligned}
E[\mathbf{z}_i \epsilon_i] &= 0 \\
E[x_i \epsilon_i] &= ?
\end{aligned}
$$

Choosing Instrumental Variables

$$
\begin{aligned}
y_i &= \beta x_i + \epsilon_i \\
x_i &= \Pi_1' \mathbf{z}_i^{(1)} + \Pi_2' \mathbf{z}_i^{(2)} + v_i
\end{aligned}
$$

$$
\begin{aligned}
E[\mathbf{z}_i^{(1)} \epsilon_i] &= 0 \\
E[\mathbf{z}_i^{(2)} \epsilon_i] &= ?
\end{aligned}
$$

FMSC Asymptotics – Local Mis-Specification



Local Mis-Specification for OLS versus TSLS

$$
\begin{aligned}
y_i &= \beta x_i + \epsilon_i \\
x_i &= \mathbf{z}_i' \boldsymbol{\pi} + v_i
\end{aligned}
$$

$$
\begin{aligned}
E[\mathbf{z}_i \epsilon_i] &= 0 \\
E[x_i \epsilon_i] &= \tau / \sqrt{n}
\end{aligned}
$$

Local Mis-Specification for Choosing IVs

$$
\begin{aligned}
y_i &= \beta x_i + \epsilon_i \\
x_i &= \Pi_1' \mathbf{z}_i^{(1)} + \Pi_2' \mathbf{z}_i^{(2)} + v_i
\end{aligned}
$$

$$
\begin{aligned}
E[\mathbf{z}_i^{(1)} \epsilon_i] &= 0 \\
E[\mathbf{z}_i^{(1)} \epsilon_i] &= \tau/\sqrt{n}
\end{aligned}
$$

Local Mis-Specification

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

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

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

(c) $\{f(Z_{ni}, \theta_0) \colon 1 \leq i \leq n, n = 1, 2, \ldots\}$ 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

$$\widehat{\theta}_S = \underset{\theta \in \Theta}{\arg\min} \ [\Xi_S f_n(\theta)]' \, \widetilde{W}_S \ [\Xi_S f_n(\theta)]$$

$$\begin{aligned}
\Xi_S &= \text{Selection Matrix (ones and zeros)} \\
\widetilde{W}_S &= \text{Weight Matrix (p.s.d.)} \\
f_n(\theta) &= \begin{bmatrix} g_n(\theta) \\ h_n(\theta) \end{bmatrix} = \begin{bmatrix} n^{-1} \sum_{i=1}^n g(Z_{ni}, \theta) \\ n^{-1} \sum_{i=1}^n h(Z_{ni}, \theta) \end{bmatrix}
\end{aligned}$$

Notation: Limit Quantities

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

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

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

Local Mis-Specification + Standard Regularity Conditions

Every candidate estimator is consistent for θ_0 and

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

$$\begin{aligned} K_S &= [F_S' W_S F_S]^{-1} F_S' W_S \\ M &= (M_g', M_h')' \\ M &\sim N(0, \Omega) \end{aligned}$$

Scalar Target Parameter μ

$$\begin{aligned}
\mu &= \mu(\theta) && \text{Z-a.s. continuous function}\\
\mu_0 &= \mu(\theta_0) && \text{true value}\\
\widehat{\mu} &= \mu(\widehat{\theta}_S) && \text{estimator}
\end{aligned}$$

Delta Method

$$\sqrt{n}\,(\widehat{\mu}_S - \mu_0) \to_d -\nabla_\theta \mu(\theta_0)' K_S \Xi_S \left(M + \left[\begin{array}{c} 0 \\ \tau \end{array} \right] \right)$$

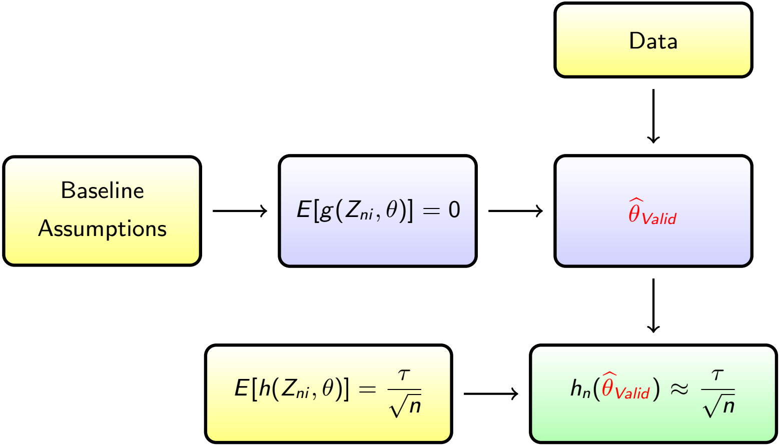
FMSC: Estimate AMSE($\widehat{\mu}_S$) and minimize over S

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

Estimating the unknowns

No consistent estimator of τ exists! (But everything else is easy)

A Plug-in Estimator of τ



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

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

$$\Psi = \left[\begin{array}{cc} -HK_v & \mathbf{I}_q \end{array} \right]$$

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

FMSC: Asymptotically Unbiased Estimator of AMSE

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

$$\widehat{B} = \widehat{\tau}\widehat{\tau}' - \widehat{\Psi}\widehat{\Omega}\widehat{\Psi}'$$

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

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 $\alpha \approx 0.16$

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

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

Limit Distribution of FMSC

$FMSC_n(S) \to_d FMSC_S$, where

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

Conservative criterion: random even in the limit.

Moment Average Estimators

$$\widehat{\mu} = \sum_{S \in \mathscr{S}} \widehat{\omega}_S \widehat{\mu}_S$$

Additional Notation

$\widehat{\mu}$	Moment-average Estimator
$\widehat{\mu}_S$	Estimator of target parameter under moment set S
$\widehat{\omega}_S$	Data-dependent weight function
$\mathscr{S}$	Collection of moment sets under consideration

Examples of Moment-Averaging Weights

Post-Moment Selection Weights

$\widehat{\omega}_S = \mathbf{1}\left\{\text{MSC}_n(S) = \min_{S' \in \mathscr{S}} \text{MSC}_n(S')\right\}$

Exponential Weights

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

Minimum-AMSE Weights...

Minimum AMSE-Averaging Estimator: OLS vs. TSLS

$$\widetilde{\beta}(\omega) = \omega \widehat{\beta}_{OLS} + (1-\omega)\widetilde{\beta}_{TSLS}$$

Under homoskedasticity:

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

Estimate by:

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

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

Limit Distribution of Moment-Average Estimators

$$\widehat{\mu} = \sum_{S \in \mathscr{S}} \widehat{\omega}_S \widehat{\mu}_S$$

(i) $\sum_{S \in \mathscr{S}} \widehat{\omega}_S = 1$ a.s.

(ii) $\widehat{\omega}(S) \rightarrow_d \varphi_S(\tau, M)$ a.s.-continuous function of τ, M and consistently-estimable constants only

$$\sqrt{n}\left(\widehat{\mu} - \mu_0\right) \rightarrow_d \Lambda(\tau)$$

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

Simulating from the Limit Experiment

Suppose τ Known, Consistent Estimators of Everything Else

1. `for` $j \in \{1, 2, \ldots, J\}$
   (i) $M_j \overset{iid}{\sim} N_{p+q}\left(0, \widehat{\Omega}\right)$
   (ii) $\Lambda_j(\tau) = -\nabla_\theta \mu(\widehat{\theta})' \left[\sum_{S \in \mathscr{S}} \widehat{\varphi}_S(M_j + \tau)\widehat{K}_S \Xi_S\right](M_j + \tau)$
2. Using $\{\Lambda_j(\tau)\}_{j=1}^J$ calculate $\widehat{a}(\tau)$, $\widehat{b}(\tau)$ such that
$$P\left[\widehat{a}(\tau) \leq \Lambda(\tau) \leq \widehat{b}(\tau)\right] = 1 - \alpha$$
3. $P\left[\widehat{\mu} - \widehat{b}(\tau)/\sqrt{n} \leq \mu_0 \leq \widehat{\mu} - \widehat{a}(\tau)/\sqrt{n}\right] \approx 1 - \alpha$

Two-step Procedure for Conservative Intervals

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

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

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

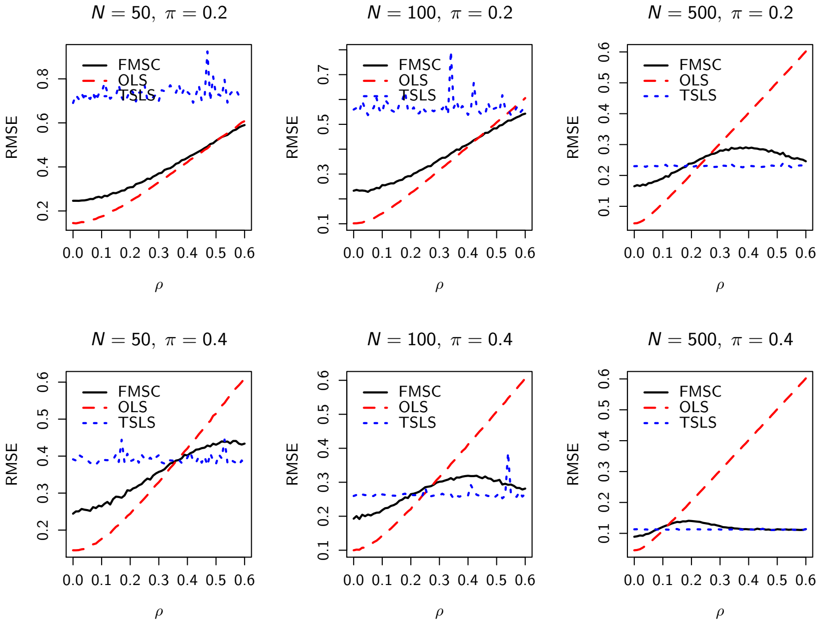
OLS versus TSLS Simulation

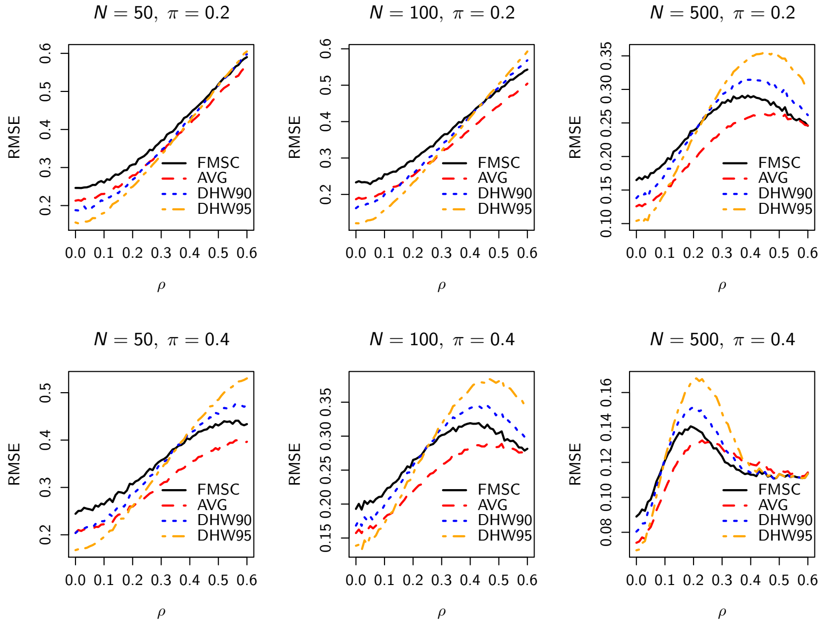
$$\begin{aligned} y_i &= 0.5x_i + \epsilon_i \\ x_i &= \pi(z_{1i} + z_{2i} + z_{3i}) + v_i \end{aligned}$$

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

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

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


$N = 50,\ \pi = 0.2$
$N = 100,\ \pi = 0.2$
$N = 500,\ \pi = 0.2$
$N = 50,\ \pi = 0.4$
$N = 100,\ \pi = 0.4$
$N = 500,\ \pi = 0.4$
FMSC
OLS
TSLS
RMSE
ρ



$N = 50, \ \pi = 0.2$
$N = 100, \ \pi = 0.2$
$N = 500, \ \pi = 0.2$
$N = 50, \ \pi = 0.4$
$N = 100, \ \pi = 0.4$
$N = 500, \ \pi = 0.4$
RMSE
ρ
FMSC
AVG
DHW90
DHW95

Choosing Instrumental Variables Simulation

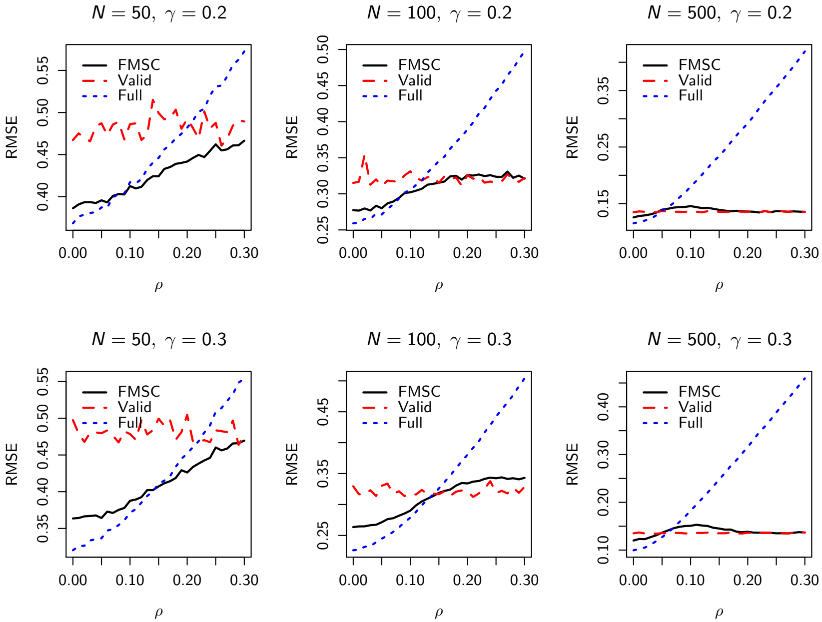
$$
\begin{aligned}
y_i &= 0.5x_i + \epsilon_i \\
x_i &= (z_{1i} + z_{2i} + z_{3i})/3 + \gamma w_i + v_i
\end{aligned}
$$

$$(\epsilon_i, v_i, w_i, z_{i1}, z_{2i}, z_{3i})' \sim \text{ iid } N(0, \mathcal{V})$$

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

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

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


$N = 50, \ \gamma = 0.2$
$N = 100, \ \gamma = 0.2$
$N = 500, \ \gamma = 0.2$
$N = 50, \ \gamma = 0.3$
$N = 100, \ \gamma = 0.3$
$N = 500, \ \gamma = 0.3$
FMSC
Valid
Full
RMSE
ρ
0.00
0.10
0.20
0.30

Alternative Moment Selection Procedures

Downward J-test

Use Full instrument set unless J-test rejects.

Andrews (1999) – GMM Moment Selection Criteria

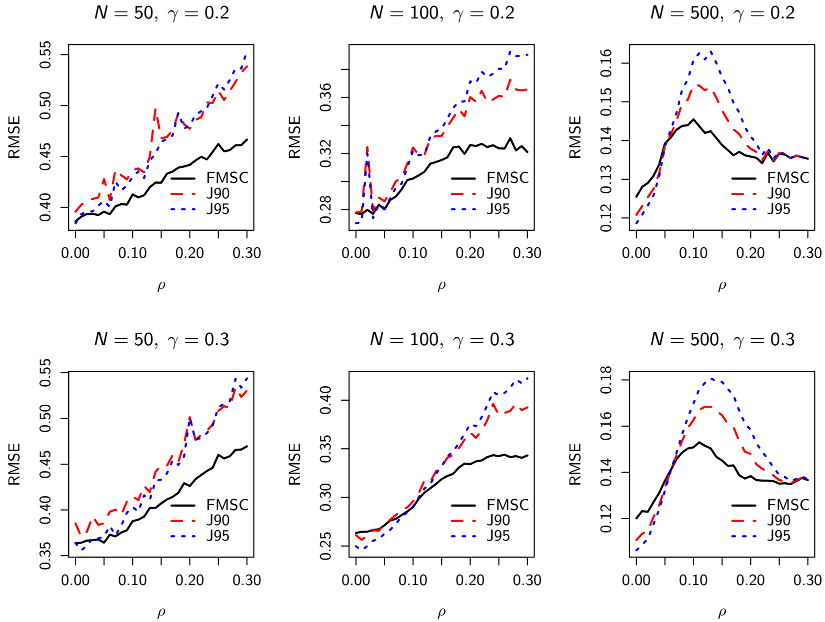
$$\text{GMM-MSC}(S) = J_n(S) - \text{Bonus}$$

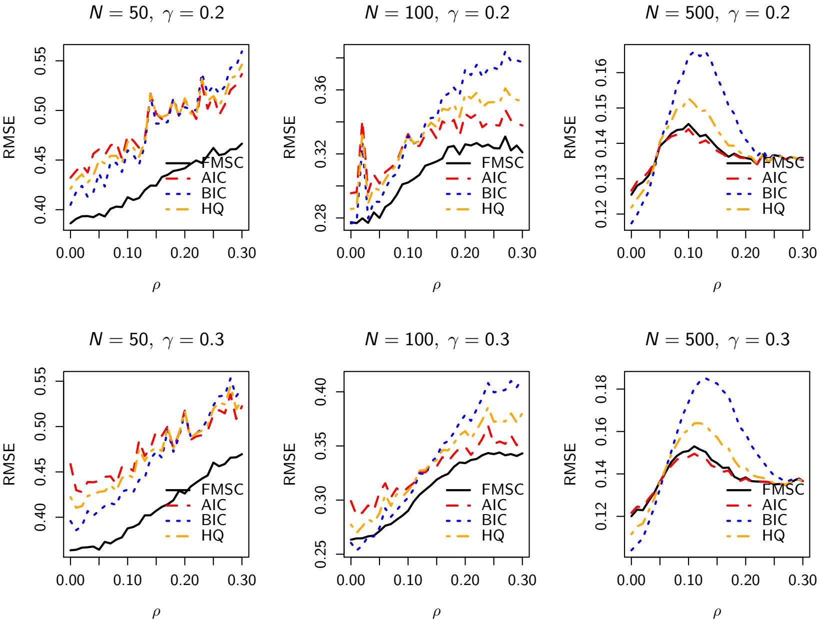
Hall & Peixe (2003) – Canonical Correlations Info. Criterion

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

Penalty/Bonus Terms

Analogies to AIC, BIC, and Hannan-Quinn


$N = 50,\ \gamma = 0.2$
$N = 100,\ \gamma = 0.2$
$N = 500,\ \gamma = 0.2$
$N = 50,\ \gamma = 0.3$
$N = 100,\ \gamma = 0.3$
$N = 500,\ \gamma = 0.3$
RMSE
ρ
FMSC
J90
J95


$N = 50,\ \gamma = 0.2$
$N = 100,\ \gamma = 0.2$
$N = 500,\ \gamma = 0.2$
$N = 50,\ \gamma = 0.3$
$N = 100,\ \gamma = 0.3$
$N = 500,\ \gamma = 0.3$
RMSE
ρ
FMSC
AIC
BIC
HQ
0.00
0.10
0.20
0.30
0.40
0.45
0.50
0.55
0.28
0.32
0.36
0.12
0.13
0.14
0.15
0.16
0.25
0.30
0.35
0.40
0.18

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
- β_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

- ln*mort* – 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*

	$\mu = malfal$			$\mu = rule$		
	FMSC	posFMSC	$\widehat{\mu}$	FMSC	posFMSC	$\widehat{\mu}$
(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)

Recall: Gauss-Markov Theorem

Linear Regression Model

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

Best Linear Unbiased Estimator

- $\text{Var}(\epsilon|X) = \sigma^2 I \Rightarrow$ then OLS has lowest variance among linear, unbiased estimators of β.
- $\text{Var}(\varepsilon|X) \neq \sigma^2 I \Rightarrow$ 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 $\widehat{\theta}$ and $\widetilde{\theta}$ be estimators of θ.

Dominance

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

Admissibility

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

Inadmissiblility

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

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

Goal

Estimate the p-vector of unknown parameters θ using X.

Maximum Likelihood Estimator $\widehat{\theta}$

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

MSE of $\widehat{\theta}$

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

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 ($\widehat{\theta}'\widehat{\theta}$ small) gives more shrinkage

MSE of James-Stein Estimator

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

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

Simplifying the Second Term

Writing Numerator as a Sum

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

For $i = 1, \ldots, p$

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

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

Combining

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

The MLE is Inadmissible when $p \geq 3$

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

- $\mathbb{E}[1/(X'X)]$ 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:

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

  - $\widehat{\beta} = \text{OLS}$, $(x)_+ = \max(x, 0)$, $\widehat{\sigma}^2 =$ usual OLS-based estimator
  - Stops us us from shrinking *past* zero to get a negative estimate for an element of β 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 $X = QR$

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

In other words, $\widehat{\beta}$ 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 β_{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 jth column of A.
  - Let $\mathbf{e}_j$ be the jth standard basis vector.
  - Inverting R is equivalent to solving $R\mathbf{a}_1 = \mathbf{e}_1$, followed by $R\mathbf{a}_2 = \mathbf{e}_2$, $\ldots$, $R\mathbf{a}_k = \mathbf{e}_k$.
- 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* $d_1, \ldots, d_r$. All other elements are zero.
- The singular values $d_1, \ldots, d_r$ 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 Λ 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 Λ 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 $d_1 \leq d_2 \leq \cdots \leq d_r$
- 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 Σ.

Optimization Problem

$$\boldsymbol{\alpha}_1 = \arg\max_{\boldsymbol{\alpha}} \ \mathrm{Var}(\boldsymbol{\alpha}'\mathbf{x}) \quad \text{subject to} \quad \boldsymbol{\alpha}'\boldsymbol{\alpha} = 1$$

First Principal Component

The linear combination $\boldsymbol{\alpha}_1'\mathbf{x}$ is the first principal component of $\mathbf{x}$.

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

Solving for $\boldsymbol{\alpha}_1$

Lagrangian

$\mathcal{L}(\boldsymbol{\alpha}_1, \lambda) = \boldsymbol{\alpha}'\Sigma\boldsymbol{\alpha} - \lambda(\boldsymbol{\alpha}'\boldsymbol{\alpha} - 1)$

First Order Condition

$2(\Sigma\boldsymbol{\alpha}_1 - \lambda\boldsymbol{\alpha}_1) = 0 \iff (\Sigma - \lambda I_p)\boldsymbol{\alpha}_1 = 0 \iff \Sigma\boldsymbol{\alpha}_1 = \lambda\boldsymbol{\alpha}_1$

Variance of 1st PC

$\boldsymbol{\alpha}_1$ is an e-vector of Σ but which one? Substituting,

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

Solution

Var. of 1st PC equals λ and this is what we want to maximize, so $\boldsymbol{\alpha}_1$ 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 $\boldsymbol{\alpha}_2'\mathbf{x}$ is uncorrelated with $\boldsymbol{\alpha}_1'\mathbf{x}$.

jth Principal Component

The linear combination $\boldsymbol{\alpha}_j'\mathbf{x}$ where $\boldsymbol{\alpha}_j$ is the e-vector corresponding to the jth largest e-value of Σ.

Sample PCA

Notation

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

SVD

$X = UDV'$, 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 jth column of V. Then,

$$\begin{aligned} \mathbf{v}_j &= \text{PC loadings for jth PC of } S \\ \mathbf{v}_j'\mathbf{x}_i &= \text{PC score for individual/time period } i \end{aligned}$$

Sample PCA

PC scores for jth 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.

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

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 $X'X = VD^2V'$

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

Lecture #9 – High-Dimensional Regression II

Ridge Regression

Principal Components Regression

LASSO

Ridge Regression – OLS with an L_2 Penalty

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

- 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 β_1 becomes β_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

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

- 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 λ (data-dependend)

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 σ^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,

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

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

Ridge Regresion via OLS with "Dummy Observations"

Ridge regression solution is identical to

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

where

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

since:

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

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:

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

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

Columns of $\sqrt{\lambda} I_p$ are linearly independent, so columns of $\widetilde{X}$ 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 $\widetilde{X} = QR$ so

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

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

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 $\underset{(n \times p)}{X}$ with rank p so OLS estimator is unique.

Economical SVD

- $\underset{(n \times p)}{X} = \underset{(n \times p)}{U} \underset{(p \times p)}{D} \underset{(p \times p)}{V'}$ 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,

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

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 $H(\lambda)$

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

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

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

Effective Degrees of Freedom for Ridge Regression

Using cyclic permutation property of trace:

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

- $\mathrm{df}(\lambda) \to 0$ as $\lambda \to \infty$
- $\mathrm{df}(\lambda) = p$ when $\lambda = 0$
- $\mathrm{df}(\lambda) < p$ when $\lambda > 0$

Comparing OLS and Ridge Predictions

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

Comparing OLS and Ridge Predictions

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

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

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

- 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: $\mathbf{z}_i = d_j \mathbf{u}_j$.
3. Throw away all but first k principal components, where $k < p$.
4. Regress $\mathbf{y}$ on $\mathbf{z_1}, \ldots, \mathbf{z}_k$.

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

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

$$\widehat{\beta}_{Ridge} = \underset{\beta}{\arg\min}\; (\mathbf{y} - X\beta)'(\mathbf{y} - X\beta) + \lambda\, ||\beta||_2^2$$

LASSO – L_1 Penalty

$$\widehat{\beta}_{Lasso} = \underset{\beta}{\arg\min}\; (\mathbf{y} - X\beta)'(\mathbf{y} - X\beta) + \lambda\, ||\beta||_1$$

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 λ and t.

Bayesian Posterior Mode

Ignoring the intercept, LASSO is the posterior model for β 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 $\text{Lap}(x|\mu, \tau) = (2\tau)^{-1} \exp\left\{-\tau^{-1}|x - \mu|\right\}$

Comparing Ridge and LASSO – Bayesian Posterior Modes

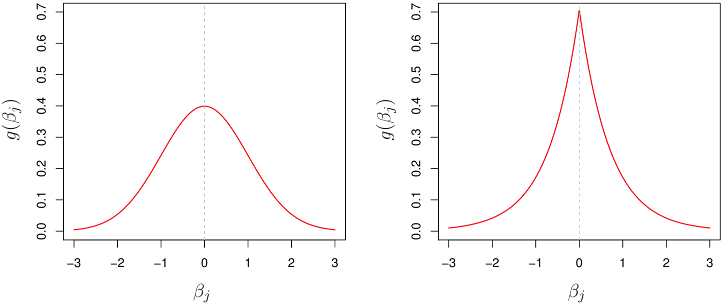


Figure: Ridge, at left, puts a normal prior on β while LASSO, at right, uses a Laplace prior, which has fatter tails and a taller peak at zero.

Comparing LASSO and Ridge – Constrained OLS

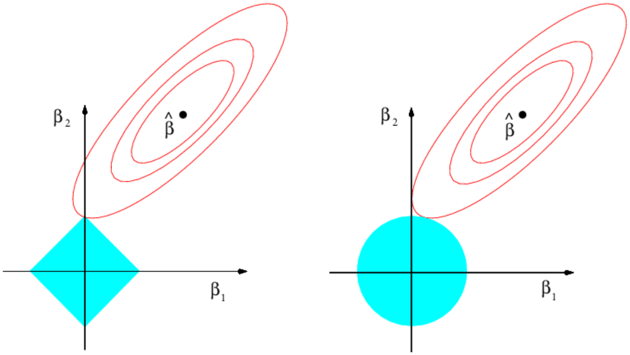


Figure: $\widehat{\beta}$ denotes the MLE and the ellipses are the contours of the likelihood. LASSO, at left, and Ridge, at right, both shrink β away from the MLE towards zero. Because of its diamond-shaped constraint set, however, LASSO favors a sparse solution while Ridge does not

No Closed-Form for LASSO!

Simple Special Case

Suppose that $X'X = I_p$

Maximum Likelihood

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

Ridge Regression

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

So what about LASSO?

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

Want to Solve

$$\widehat{\boldsymbol{\beta}}_{LASSO} = \underset{\boldsymbol{\beta}}{\arg\min}\,(\mathbf{y} - X\boldsymbol{\beta})'(\mathbf{y} - X\boldsymbol{\beta}) + \lambda\,||\boldsymbol{\beta}||_1$$

Expand First Term

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

Hence

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

LASSO when $X'X = I_p$

Preceding Slide

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

Key Simplification

Equivalent to solving j independent optimization problems:

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

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

LASSO when $X'X = I_p$

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

Optimization problem becomes

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

Interior solution:

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

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

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

LASSO when $X'X = I_p$

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

Optimization problem becomes

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

Interior solution:

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

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

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

Ridge versus LASSO when $X'X = I_p$

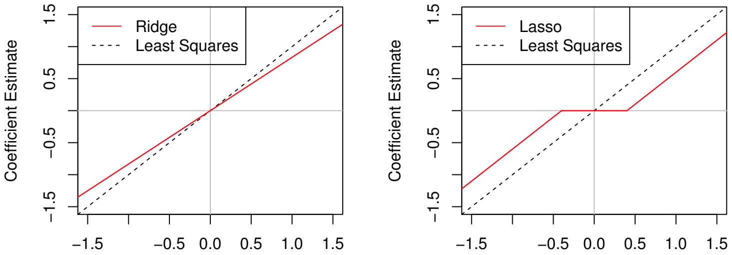


Figure: Horizontal axis in each plot is MLE

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

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

Calculating LASSO – The Shooting Algorithm

Cyclic Coordinate Descent

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

Result: LASSO Solution

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

repeat

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

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

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

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

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

  end

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

Lecture #10 – Factor Models etc.

Overview of lots of stuff since we're short on time!

Survey Articles on Factor Models

Stock & Watson (2010)

Best general overview of factor models and applications.

Bai & Ng (2008)

Comprehensive review of large-sample results for high-dimensional factor models estimated via PCA.

Stock & Watson (2006)

Handbook chapter on forecasting with many predictors. One section is devoted to dynamic factor models.

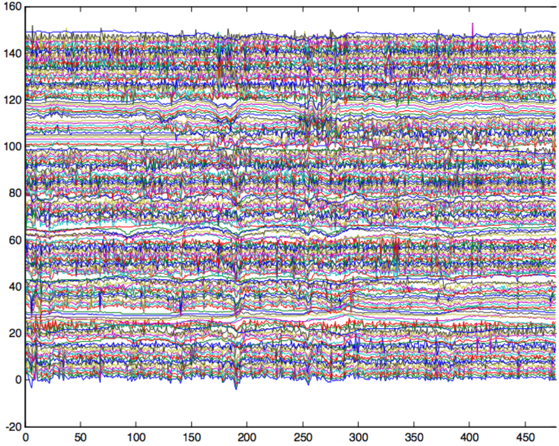
Breitung & Eickmeyer (2006)

Brief overview with an application to Euro-area business cycles.

The Basic Idea

We're interested in settings with a large number of time series N and a comparable number of time periods T.

Example: Stock and Watson Dataset



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

Why Factor Models?

1. Factors could be intrinsically interesting if they arise from a theoretical model (e.g. Financial Economics)
2. Many variables without running out of degrees of freedom
   - More information could improve forecasts/macro analysis
   - Mimic central banks "looking at everything"
3. Eliminate measurement error and idiosyncratic shocks to provide more reliable information for policy
4. "Remain Agnostic about the Structure of the Economy"
   - Advantages over SVARs: don't have to choose variables to control degrees of freedom, and can allow fewer underlying shocks than variables.

Classical Factor Analysis Model

Assume that X_t has been de-meaned...

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

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

$\Lambda =$ matrix of factor loadings

$\Psi =$ diagonal matrix of idiosyncratic variances.

Adding Time-Dependence

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

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

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

Terminology

Static	X_t depends only on F_t
Dynamic	X_t depends on lags of F_t as well
Exact	Ψ is diagonal and ϵ_t independent over time
Approximate	Some cross-sectional & temporal dependence in ϵ_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. The difference between "static" and "dynamic" is unclear
   - 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. Not really possible to allow cross-sectional dependence in ϵ_t
   - Unless the off-diagonal elements of Ψ 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 Ψ are negligible
   - Similarly, time series dependence in ϵ_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
   - Just carry out PCA on X and ignore the time-series element
   - The first r PCs are our estimates $\widehat{F}_t$
   - 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 $F = (F_1, \ldots, F_T)'$
- $\Lambda'\Lambda = \text{diag}(\mu_1, \ldots, \mu_r)$ where $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_r$

Assumption I

Factors are *pervasive*: $\Lambda'\Lambda/N \to D_\Lambda$ 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

If we average over the cross-section, the contribution from the factors persists and the contribution from the idiosyncratic terms disappears as $N \to \infty$.

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 $N, T \to \infty$.

Doz, Giannone & Reichlin (2012)

The arguments for the PCA approach...

- Consistent estimation of factors under very weak assumptions
- MLE is computationally infeasible for large N

...may be somewhat exaggerated.

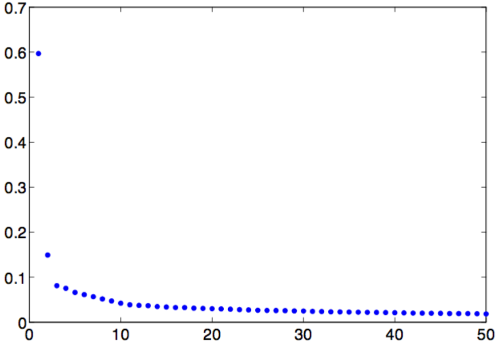
- EM-algorithm + Kalman Filter is *very efficient* – complexity depends on number of *factors*, not number of series
- Treat exact, static factor model (the one I wrote out) as a mis-specified *approximating model* (Quasi-MLE)
- Identical large-sample results as PC under similar assumptions, but better finite-sample properties and temporal smoothing

Choosing the Number of Factors

If we use Likelihood-based or Bayesian estimation, we could try to resort to the familiar tools from earlier in the semester. There are a lot of parameters in factor models, however, so the asymptotic approximations (I'm looking at you, AIC) could be poor.

Choosing the Number of Factors – Scree Plot

If we use PC estimation, we can look a 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:

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

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.

What Can We Do with Factors?

Among other possibilities:

1. Use them to construct Forecasts
2. Use them as Instrumental Variables
3. Use them to “Augment” a VAR

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).

Main References

Stock & Watson (2006) – “Forecasting with Many Predictors”

Overview of high-dimesional forecasting with a review of forecast combination, factor models, and Bayesian approaches.

Ng (2013) – “Variable Selection in Predictive Regressions”

Reviews and relates a number of shrinkage & selection methods.

Stock & Watson (2012)

Examines a wide range of shrinkage procedures to see if they can improve on diffusion index forecasts.

Kim & Nelson (2013)

“Horse Race” of various factor and shrinkage methods for forecasting.

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 $t = 1, \ldots, T$.

Assumption

Static representation of Dynamic Factor Model:

$$\begin{aligned} y_t &= \beta' F_t + \gamma(L) y_t + \epsilon_{t+1} \\ X_t &= \Lambda F_t + e_t \end{aligned}$$

"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 $\widehat{F}_t$
    - Missing observations/Mixed-frequency: EM-algorithm
3. Fit the Forecasting Regression
    - Regress y_t on a constant and lags of $\widehat{F}_t$ 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 (ϵ, e, F) plus a "rank condition" on Λ imply that the MSE of the procedure on the previous slide converges to that of the infeasible optimal procedure, provided that $N, T \to \infty$.

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

Factors as Instruments – Bai & Ng (2010)

Endogenous Regressors x_t

$$y_t = x_t'\beta + \epsilon_t \qquad E[x_t\epsilon_t] \neq 0$$

Unobserved Variables F_t are Strong IVs

$$\underset{(k\times 1)}{x_t} = \Psi' \underset{(r\times 1)}{F_t} + u_t \qquad E[F_t\epsilon_t] = 0$$

Observe Large Panel $(z_{1t}, \ldots, z_{Nt})$

$$z_{it} = \lambda_i' F_t + e_{it}$$

Factors as Instruments – Bai & Ng (2010)

$$\boxed{y_t = x_t'\beta + \epsilon_t, \qquad x_t = \Psi' F_t + u_t, \qquad z_{it} = \lambda_i' F_t + e_{it}}$$

Procedure

1. Calculate the PCs of Z
2. Calculate $\widetilde{F}_t$ using the first r PCs of Z
3. Use $\widetilde{F}_t$ in place of F_t for IV estimation

Main Result

Under certain assumptions, as $(N, T) \to \infty$ "estimation and inference can proceed as though F_t were known." The resulting estimator is consistent and asymptotically normal.

Factors as Instruments – Bai & Ng (2010)

Why Might This be Helpful?

1. Avoid many instruments bias
2. Avoid bias from irrelevant instruments
3. Allow more observed instruments z_{it} than sample size T
4. Provided that $\sqrt{T}/N \to 0$, all of the observed instruments z_{it} can be *endogenous* as long as F_t is exogenous

FAVARs – Bernanke, Boivin & Eliasz (2005)

Two Problems with Structural VARs

1. Number of parameters is *quadratic* in the number of variables. Unrestricted VAR infeasible unless T is large relative to N.
   - You've studied one solution to this problem already this semester: Bayesian Estimation with informative priors
2. To keep estimation tractable we typically use a small number of variables, but then the VAR innovations "might not span the space of structural shocks."

FAVARs – Bernanke, Boivin & Eliasz (2005)

Factor-Augmented VAR Model

$$\begin{bmatrix} Y_t \\ F_t \end{bmatrix} = \Phi(L) \begin{bmatrix} F_{t-1} \\ Y_{t-1} \end{bmatrix} + v_t$$

$$X_t = \Lambda^f F_t + \Lambda^y Y_t + e_t$$

$\underset{(M\times 1)}{Y_t}$ = observable variables that "drive dynamics of the economy"

$\underset{(K\times 1)}{F_t}$ = Small # of unobserved factors: "additional information"

$\underset{(N\times 1)}{X_t}$ = Large # of observed "informational time series"

FAVARs – Bernanke, Boivin & Eliasz (2005)

$$\boxed{\begin{bmatrix} Y_t \\ F_t \end{bmatrix} = \Phi(L) \begin{bmatrix} F_{t-1} \\ Y_{t-1} \end{bmatrix} + v_t \qquad X_t = \Lambda^f F_t + \Lambda^y Y_t + e_t}$$

Consider Two Estimation Procedures

1. Two-step Procedure:
   - Estimate space spanned by factors using first $K + M$ PCs of X
   - Estimate VAR with $\widehat{F}_t$ in place of F_t
2. Full Bayes (Gibbs Sampler)

Empirical Application

Additional information contained in FVAR is "important to properly identify the monetary transmission mechanism."

What about Ridge and Lasso?

Basic Idea

Diffusion index forecasts are really just PCR. Why not try Ridge or Lasso with all predictors rather than estimating factors?

De Mol, Giannone & Reichlin (2008)

- Compare PCA-based factor forecasts to Ridge and Lasso
- In a small out-of-sample experiment, Ridge and Lasso with appropriate penalty parameters give results comparable to diffusion index.
- Analyze asymptotics of Ridge under assumptions typically used to justify PCA

Other Ways of Extracting Factors

Sparse PCA

Add a Lasso-type penalty to the “regression” formulation of PCA: encourage the factors to load on small number of variables.

Independent Components Analysis (ICA)

Extract factors that maximize non-Gaussianity

Both of these are considered in Kim & Swanson (2014) and seem to work very well when combined with second-stage shrinkage.

To Target or Not to Target?

Problem with PCA and Friends

Completely ignores Y in constructing the factors! Should we take the forecast target into account when extracting factors?

Some References

- Bai & Ng (2008) – Forecasting Economic Time Series Using Targeted Predictors
- Kelly & Pruitt (2012) – The Three-pass Regression Filter

Partial Least Squares (PLS)

As an Optimization Problem

Construct a sequence of linear combinations of X that solve

$$\max_{\alpha} Corr^2(\mathbf{y}, X\alpha) Var(X\alpha)$$

subject to $||\alpha|| = 1$ and the constraint that each PLS "factor" is orthogonal to the preceding ones.

As a Probabilistic Model

"Shared" factor F_t and X-specific factor Z_t

$$\begin{aligned} Y_t &= \mu_Y + \Lambda_Y F_t + \epsilon_t \\ X_t &= \mu_X + \Lambda_X F_t + \Pi Z_t + u_t \end{aligned}$$

where $F_t \perp Z_t$

Bootstrap Aggregation – “Bagging”

Bagging Algorithm

1. Make a bootstrap draw
2. Carry out selection/shrinkage/estimation using boostrap data
3. Use estimated parameters from to construct a forecast $\widehat{y}_{T+h}^{(b)}$
4. Repeat for $b = 1, \ldots, B$
5. Average to get “Bagged” Forecast: $\widehat{y}_{T+h}^{(Bag)} = \frac{1}{B}\sum_{b=1}^{B} \widehat{y}_{T+h}^{(b)}$

Details

- If the data are dependent, need block bootstrap.
- In step 3, we forecast using the *parameters* estimated from the bootstrap data but the *predictors* from the *real* dataset.

Bootstrap Aggregation – "Bagging"

Why Bagging?

- Aims to reduce the forecast error of "unstable" procedures such as variable selection of Lasso, by reducing their variance.
- Completely portable: you can bag *anything* provided you have an appropriate way to carry out the bootstrap.
- May provide a way of attacking the problem of inference post-model selection. See Efron (JASA, *Forthcoming*) "Estimation and Accuracy after Model Selection"

Bagging in Economics

Inoue & Killian (2008, JASA)

Compares performance of bagged "pre-test" estimator (variable selection via a t-test) to other methods of forecasting US Inflation. Bagging is carried out via a block bootstrap.

Stock & Watson (2012)

Among other shrinkage procedures, they consider a large-sample approximation to bagging pre-test estimators that doesn't require making bootstrap draws.

Other Papers That Use Bagging

- Hillebrand & Medeiros (2010): Realized Volatility Forecasts
- Hillebrand et al (2012): Forecasting the Equity Premium
- Kim and Swanson (2013)

Boosting

Ensemble Methods

Machine learning term for "non-Bayesian model averaging"

What is Boosting?

- Combine large number of "weak learners" (i.e. crappy predictive models) so that the *ensemble* predicts well.
- Explicitly designed around predictive loss
- Arbitrarily improve in-sample fit of arbitrarily the weak learners!

Book-Length Treatment

Shapire & Freund (2012) – *Boosting: Foundations and Algorithms*

Boosting

Bai & Ng (2009) – Boosting Diffusion Indices

Use boosting to select which lags of factors to include in a forecasting regression estimated following PCA.

Buchen & Wohlrabe (2011) – Is Boosting a Viable Alternative?

Boosting performs well compared to other methods in the example from the 2006 Stock & Watson Handbook Chapter.

Ng (2014) – Boosting Recessions