Econ 722 – Advanced Econometrics IV

Francis J. DiTraglia

University of Pennsylvania

Lecture #8 – High-Dimensional Regression II

LASSO

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

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

LASSO – L_1 Penalty

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

Other Ways of Thinking about LASSO

Constrained Optimization

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

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

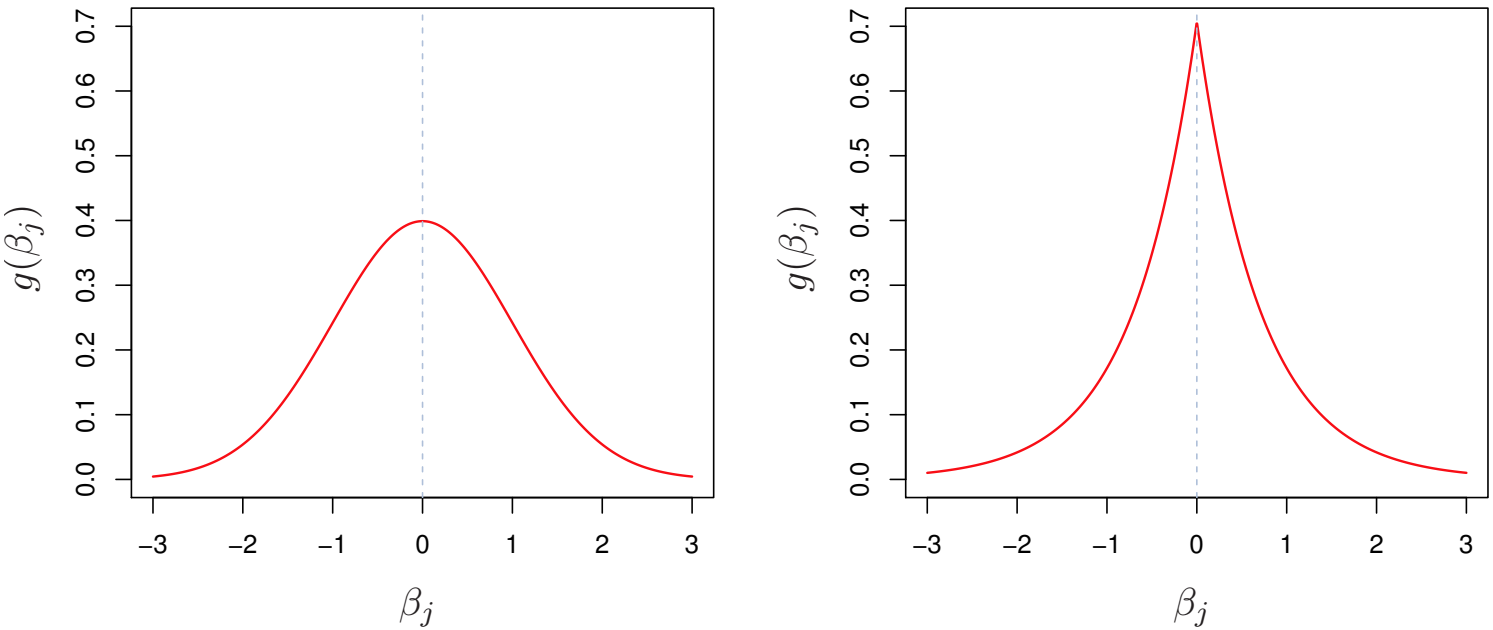


Figure: Ridge, at left, puts a normal prior on β while LASSO, at right, uses a Laplace prior, which has fatter tails and a taller peak at zero.

Comparing LASSO and Ridge – Constrained OLS

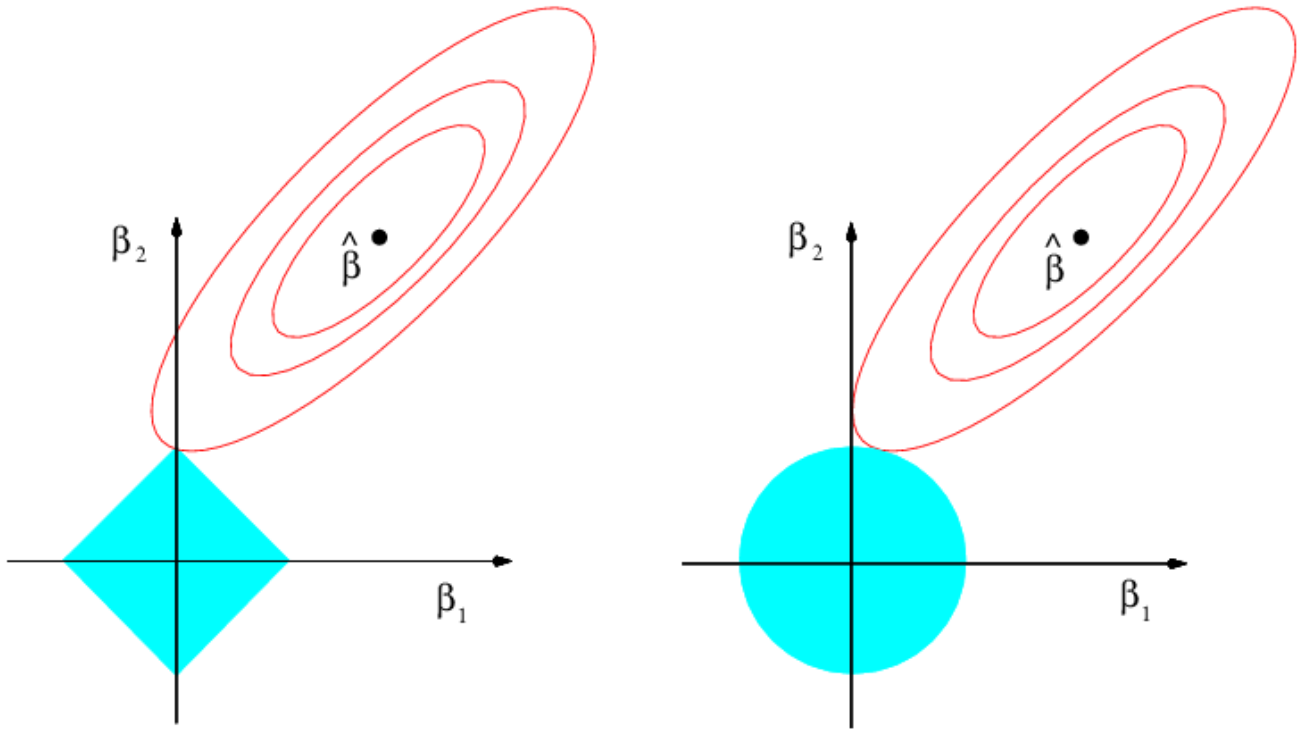


Figure: $\widehat{\beta}$ denotes the MLE and the ellipses are the contours of the likelihood. LASSO, at left, and Ridge, at right, both shrink β away from the MLE towards zero. Because of its diamond-shaped constraint set, however, LASSO favors a sparse solution while Ridge does not

No Closed-Form for LASSO!

Simple Special Case

Suppose that $X'X = I_p$

Maximum Likelihood

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

Ridge Regression

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

So what about LASSO?

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

Want to Solve

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

Expand First Term

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

Hence

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

LASSO when $X'X = I_p$

Preceding Slide

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

Key Simplification

Equivalent to solving j independent optimization problems:

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

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

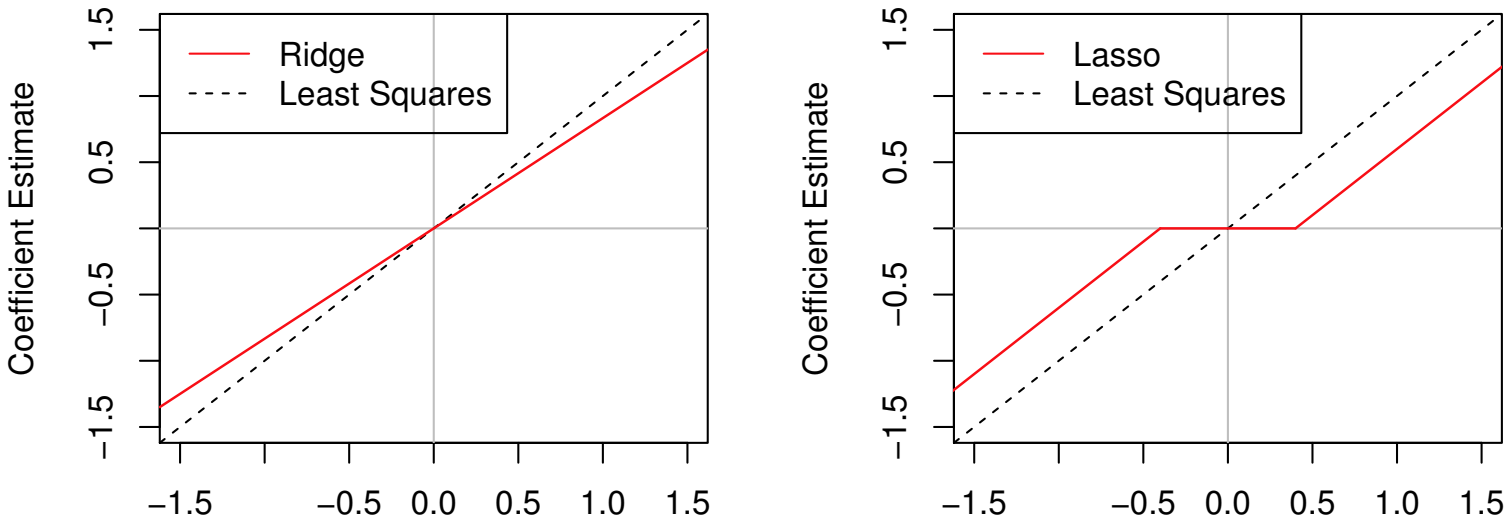


Figure: Horizontal axis in each plot is MLE

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

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

Calculating LASSO – The Shooting Algorithm

Cyclic Coordinate Descent

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

Result: LASSO Solution

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

repeat

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

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

    $a_j \leftarrow 2\sum_i x_{ij}^2$

    $c_j \leftarrow 2\sum_i x_{ij}(y_i - \mathbf{x}_i'\beta + \beta_j x_{ij})$

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

  end

until $|\beta - \beta^{prev}| < \varepsilon$;

Coordinate Updates in the Shooting Algorithm

$$\frac{\partial}{\partial \beta_j} (\mathbf{y} - X\boldsymbol{\beta})' (\mathbf{y} - X\boldsymbol{\beta}) = a_j \beta_j - c_j$$

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

$$c_j \equiv 2 \sum_{i=1}^{n} x_{ij} \Big(\underbrace{y_i - \mathbf{x}_i' \boldsymbol{\beta} + \beta_j x_{ij}}_{\text{Residual excluding } x_{ij}} \Big)$$

$$\beta_j^{\text{New}} = \begin{cases} (c_j + \lambda)/a_j, & c_j < -\lambda \\ 0, & c_j \in [-\lambda, \lambda] \\ (c_j - \lambda)/a_j, & c_j > \lambda \end{cases}$$

Prediction Error of LASSO

Punchline

With the appropriate choice of λ, Lasso can make very good predictions even when p is much larger than n, so long as $\sum_{j=1}^{p} |\beta_j|$ is small.

Sparsity?

One way to have small $\sum_{j=1}^{p} |\beta_j|$ is if $\boldsymbol{\beta}$ is *sparse*, i.e. $\beta_j = 0$ for most j, but sparsity is not required.

We'll look at a simple example...

Prediction Error of LASSO: Simple Example

Suppose that:

- X and $\mathbf{y}$ are centered
- X is fixed and scaled so that $\mathbf{x}_j'\mathbf{x}_j = n$
- $\mathbf{y} = X\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 I)$.
- $\lambda = c\sigma\sqrt{\log(p)/n}$ where c is a constant

Theorem

Let $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2n} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}_0\|_1$. Then,

$$\mathbb{P}\left(\frac{1}{n}\left\|X\boldsymbol{\beta}_{\mathbf{0}} - X\widehat{\boldsymbol{\beta}}\right\|_2^2 \leq 4\lambda \|\boldsymbol{\beta}_0\|_1\right) \geq 1 - p^{-(c^2/2-1)}$$

What Does This Mean?

$$\boxed{\mathbb{P}\left(\frac{1}{n}\left\|X\boldsymbol{\beta_0}-X\widehat{\boldsymbol{\beta}}\right\|_2^2 \leq 4\lambda\left\|\boldsymbol{\beta}_0\right\|_1\right) \geq 1-p^{-(c^2/2-1)}}$$

Notation

$\|\mathbf{z}\|_2^2 \equiv \mathbf{z}'\mathbf{z}, \quad \|\boldsymbol{\alpha}\|_1 \equiv \sum_{j=1}^p |\alpha_j|$

Convenient Scaling

Divide RSS by $2n$: $\quad \widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2n}\|\mathbf{y}-X\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}_0\|_1$

Prediction Error Comparison

Optimal: $\boldsymbol{\varepsilon} = \mathbf{y} - X\boldsymbol{\beta}_0 \quad$ Lasso: $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - X\widehat{\boldsymbol{\beta}}$

$$\frac{1}{n}\|\widehat{\boldsymbol{\varepsilon}}-\boldsymbol{\varepsilon}\|_2^2 = \frac{1}{n}\left\|(\mathbf{y}-X\widehat{\boldsymbol{\beta}})-(\mathbf{y}-X\boldsymbol{\beta_0})\right\|_2^2 = \frac{1}{n}\left\|X\boldsymbol{\beta_0}-X\widehat{\boldsymbol{\beta}}\right\|_2^2$$

What Does This Mean?

$$\boxed{\mathbb{P}\left(\frac{1}{n}\left\|X\boldsymbol{\beta_0} - X\widehat{\boldsymbol{\beta}}\right\|_2^2 \leq 4\lambda\left\|\boldsymbol{\beta}_0\right\|_1\right) \geq 1 - p^{-(c^2/2-1)}}$$

Recall

$\lambda = c\sigma\sqrt{\log(p)/n}, \quad \varepsilon \sim N(0, \sigma^2 I)$

We choose c

Larger $c \implies$ higher probability that the bound obtains:

$$c = 2 \implies 1 - p^{-(c^2/2-1)} = 1 - 1/p$$

$$c = 3 \implies 1 - p^{-(c^2/2-1)} = 1 - p^{-7/2}$$

$$c = 4 \implies 1 - p^{-(c^2/2-1)} = 1 - p^{-7}$$

What Does This Mean?

$$\boxed{\mathbb{P}\left(\frac{1}{n}\left\|X\boldsymbol{\beta_0} - X\widehat{\boldsymbol{\beta}}\right\|_2^2 \leq 4\lambda\left\|\boldsymbol{\beta}_0\right\|_1\right) \geq 1 - p^{-(c^2/2-1)}}$$

Recall

$\lambda = c\sigma\sqrt{\log(p)/n}, \quad \varepsilon \sim N(0, \sigma^2 I)$

We choose c

Larger $c \implies$ looser bound:

$$c = 2 \implies 4\lambda\left\|\boldsymbol{\beta}_0\right\|_1 = 8\sigma\sqrt{\log(p)/n} \times \left\|\boldsymbol{\beta}_0\right\|_1$$

$$c = 3 \implies 4\lambda\left\|\boldsymbol{\beta}_0\right\|_1 = 12\sigma\sqrt{\log(p)/n} \times \left\|\boldsymbol{\beta}_0\right\|_1$$

$$c = 4 \implies 4\lambda\left\|\boldsymbol{\beta}_0\right\|_1 = 16\sigma\sqrt{\log(p)/n} \times \left\|\boldsymbol{\beta}_0\right\|_1$$

We can allow $p \gg n$ provided $\|\beta\|_1$ is small

$$\mathbb{P}\left(\frac{1}{n}\left\|X\beta_0 - X\widehat{\beta}\right\|_2^2 \leq 4\lambda \|\beta_0\|_1\right) \geq 1 - p^{-(c^2/2-1)}$$

Recall

$\lambda = c\sigma\sqrt{\log(p)/n}, \quad \varepsilon \sim N(0, \sigma^2 I)$

p	n	$\sqrt{\log(p)/n}$
100	100	0.21
1000	1000	0.08
1000	100	0.26
10000	1000	0.10
10000	100	0.30
100000	1000	0.11

Lecture #9 – High-Dimensional Regression III

Principal Component Analysis (PCA)

Principal Components Regression

Comparing OLS, Ridge, and PCR

Overview of Factor Models

Choosing the Number of Factors

Diffusion Index Forecasting

Principal Component Analysis (PCA)

Notation

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

Optimization Problem

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

First Principal Component

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

The random vector $\mathbf{x}$ has maximal variation in the direction α_1.

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 $X'X = VD^2V'$

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

Principal Components Regression (PCR)

1. Start with centered X and $\mathbf{y}$.
2. SVD of $X \implies$ PC scores: $\mathbf{z}_j = X\mathbf{v}_j = d_j\mathbf{u}_j$.
3. Regress $\mathbf{y}$ on $[\ \mathbf{z}_1 \quad \dots \quad \mathbf{z}_m\ ]$ where $m < p$.

$$\widehat{\mathbf{y}}_{\text{PCR}}(m) = \sum_{j=1}^{m} \mathbf{z}_j \widehat{\theta}_j, \quad \widehat{\theta}_j = \frac{\mathbf{z}_j'\mathbf{y}}{\mathbf{z}_j'\mathbf{z}_j} \quad \text{(PCs orthogonal)}$$

Standardizing X

Because PCR is not scale invariant, it is common to standardize X.

This amounts to PCA performed on a correlation matrix.

Comparing PCR, OLS and Ridge Predictions

Assumption

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

SVD

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

Ridge Predictions

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

Relating OLS and Ridge to PCR

Recall: U is Orthonormal

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

Substituting

$$\widehat{\mathbf{y}}_{\text{Ridge}}(\lambda) = \sum_{j=1}^{m}\left(\frac{d_j^2}{d_j^2+\lambda}\right)\mathbf{u}_j\mathbf{u}_j'\mathbf{y} = \sum_{j=1}^{m}\left(\frac{d_j^2}{d_j^2+\lambda}\right)\mathbf{z}_j\widehat{\theta}_j$$

$$\widehat{\mathbf{y}}_{\text{OLS}} = \widehat{\mathbf{y}}_{\text{Ridge}}(0) = \sum_{j=1}^{p}\mathbf{z}_j\widehat{\theta}_j$$

Comparing PCR, OLS, and Ridge Predictions

$$\widehat{\mathbf{y}}_{\text{PCR}}(m) = \sum_{j=1}^{m} \mathbf{z}_j \widehat{\theta}_j, \quad \widehat{\mathbf{y}}_{\text{OLS}} = \sum_{j=1}^{p} \mathbf{z}_j \widehat{\theta}_j, \quad \widehat{\mathbf{y}}_{\text{Ridge}}(\lambda) = \sum_{j=1}^{m} \left(\frac{d_j^2}{d_j^2 + \lambda} \right) \mathbf{z}_j \widehat{\theta}_j$$

- $\mathbf{z}_j$ is the jth sample PC
- d_j^2/n is the variance of the jth sample PC
- Ridge regresses y on sample PCs but shrinks predictions towards zero: higher variance PCs are shrunk less.
- PCR truncates the PCs with the smallest variance.
- OLS neither shrinks nor truncates: is uses all the PCs.

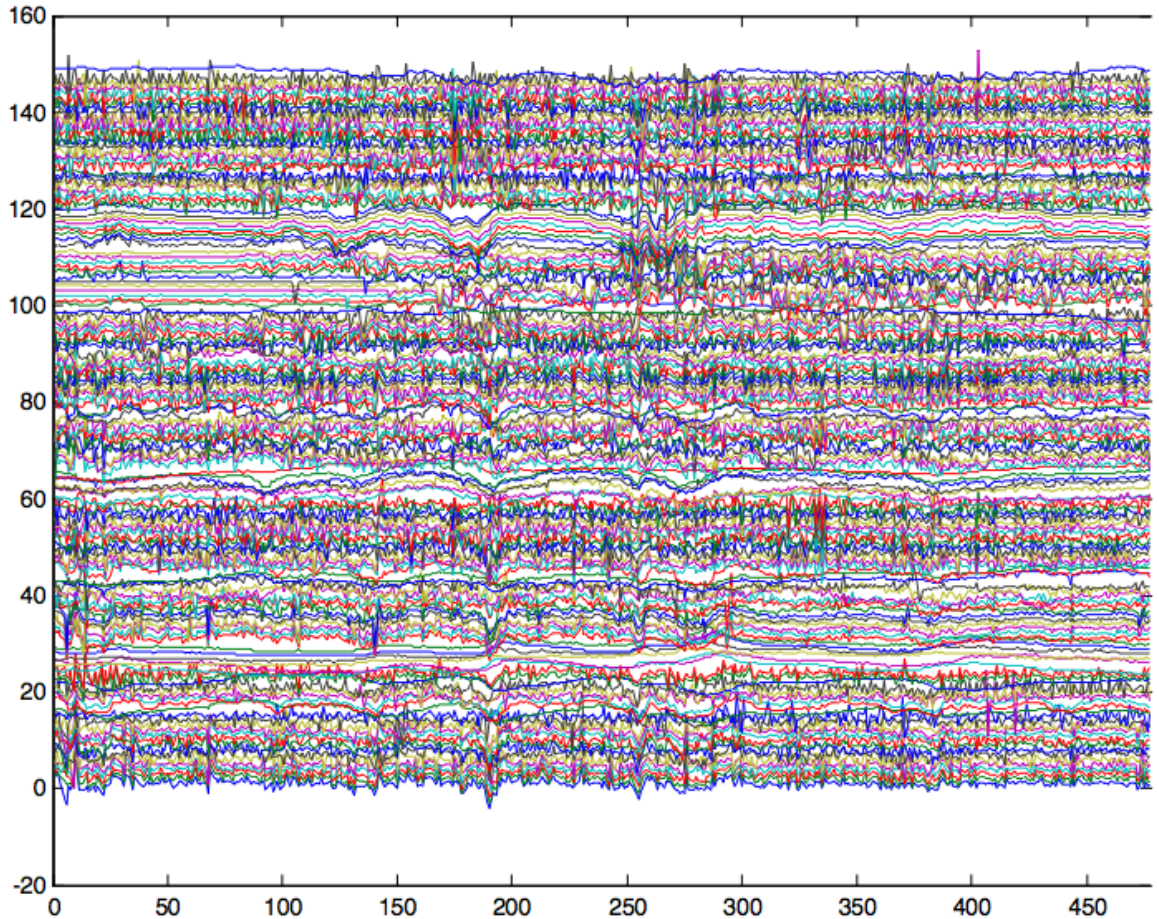
The Basic Idea

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

Survey Articles

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

Example: Stock and Watson Dataset



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

Classical Factor Analysis Model

Assume that X_t has been de-meaned...

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

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

Λ = matrix of factor loadings

Ψ = diagonal matrix of idiosyncratic variances.

Adding Time-Dependence

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

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

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

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. Are "static" and "dynamic" really different?
   - Can write dynamic model as a static one with more factors
   - Static representation involves "different" factors, but we may not care: are the factors "real" or just a data summary?
2. Can we *really* allow for cross-sectional dependence?
   - Unless the off-diagonal elements of 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 via PCA
   - PCA on the $(T \times N)$ matrix X, ignoring time dependence.
   - The $(r \times 1)$ vector $\widehat{F}_t$ of PC scores associated with the first r PCs are our estimate of F_t
   - Essentially treats F_t as an r-dimensional *parameter* to be estimated from an N-dimensional observation X_t

Estimation by PCA

PCA Normalization

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

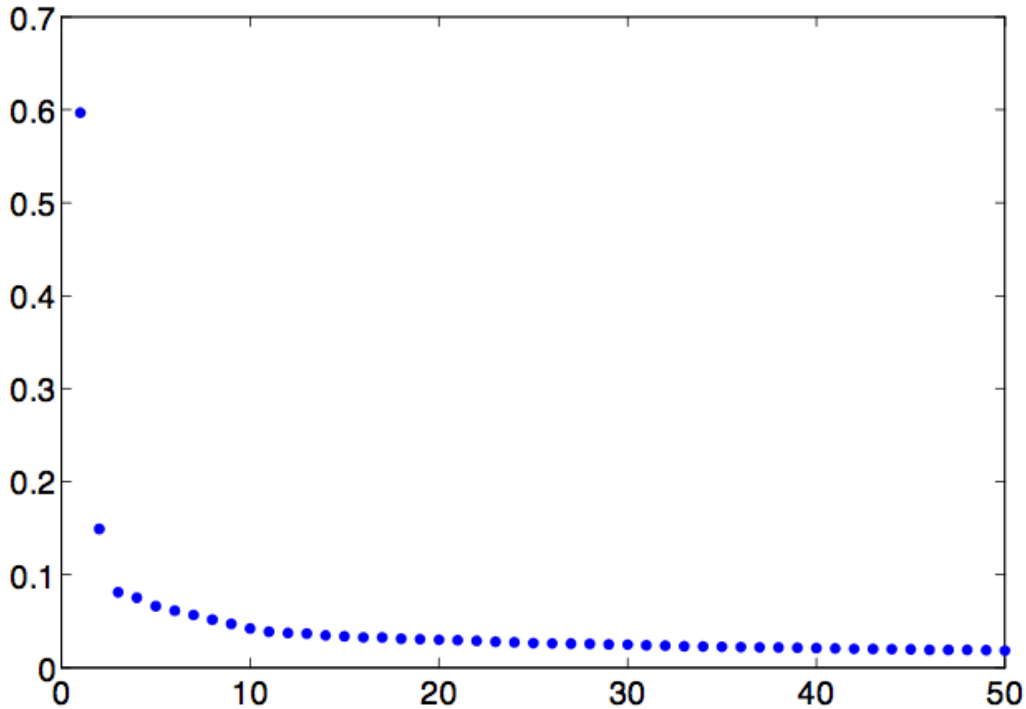
Average over the cross-section $\implies$ contribution from the factors persists while contribution from the idiosyncratic terms disappears as $N \to \infty$.

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 $N, T \to \infty$.

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.

Some Special Problems in High-dimensional Forecasting

Estimation Uncertainty

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

Best Subsets Infeasible

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

Noise Accumulation

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

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

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

Basic Setup

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

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

Lecture #10 – Selective Inference

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

How Statistics is Done In Reality

Step 1: Selection – Decide what questions to ask.

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

Step 2: Inference – Answer the Questions.

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

Problem – "Data-snooping"

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

Simple Example: "File Drawer Problem"

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

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

Two Questions

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

Simple Example: "File Drawer Problem"

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

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

Simple Example: "File Drawer Problem"

Conditional vs. Unconditional Type I Error Rates

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

Idea for Post-Selection Inference

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

Simple Example: "File Drawer Problem"

Conditional Type I Error Rate

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

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

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

$$c \approx 2.41$$

Notice:

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

Selective Inference vs. Sample-Splitting

Classical Inference

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

Selective Inference

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

Sample-Splitting

Use different datasets to choose (M, H_0) and carry out inference: $\mathbb{P}_{M,H_0}\big(\text{reject } H_0 | \{M, H_0 \text{ selected}\}\big) = \mathbb{P}_{M,H_0}(\text{reject } H_0)$.

Selective Inference in Exponential Family Models

Questions

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

Fithian, Sun & Taylor (2017)

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

A Prototype Example of Selective Inference

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

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

All Calculations are Under a Given Model M

Key Idea

Under whichever model M ends up being selected, there is a joint normal distribution for Y_β and Y_δ *without* conditioning on A.

WLOG unit variances, ρ known

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

As long as we can consistently estimate the variances of Y_β and Y_δ along with their covariance, this is not a problem.

Selective Inference in a Bivariate Normal Example

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

Two Cases

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

Problem

If δ were known, we could directly calculate how conditioning on A affects the distribution of Y_β, but δ is unknown!

Solution

Condition on a sufficient statistic for δ.

Conditioning on a Sufficient Statistic

Theorem

If U is a sufficient statistic for δ, then the joint distribution of (Y_β, Y_δ) given U does not depend on δ.

In Our Example

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

Straightforward Calculation

$$\left[\begin{array}{c} Y_\beta \\ Y_\delta \end{array} \right] \Bigg| (U = u) = \left[\begin{array}{c} \beta + Z \\ u + \rho(\beta + Z) \end{array} \right], \quad Z \sim \mathsf{N}(0,1)$$

Notice that this is a singular normal distribution

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

$$\left[\begin{array}{c} Y_\beta \\ Y_\delta \end{array}\right] \Bigg| (U = u) = \left[\begin{array}{c} \beta + Z \\ u + \rho(\beta + Z) \end{array}\right], \quad Z \sim \mathsf{N}(0, 1)$$

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

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

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

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

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

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

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

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

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

where

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

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

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

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

Inverting the CDF from the preceding slide:

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

where

$$\begin{aligned} p^* &\equiv \Phi\left(\ell(u) - \beta_0\right) / P_D(A) \\ P_D(A) &\equiv 1 - \Phi\left(r(u) - \beta_0\right) + \Phi\left(\ell(u) - \beta_0\right) \\ \ell(u) &\equiv (-c - u)/\rho \\ r(u) &\equiv (c - u)/\rho \end{aligned}$$

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

$$\left[\begin{array}{c} Y_\beta \\ Y_\delta \end{array}\right] \Bigg| (U = u) = \left[\begin{array}{c} \beta + Z \\ u + \rho(\beta + Z) \end{array}\right], \quad Z \sim \mathsf{N}(0,1)$$

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

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

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

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

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

where

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

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

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

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

Inverting the CDF from the preceding slide:

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

where:

$$\begin{aligned} P_D(A^c) &\equiv \Phi\left(r(u) - \beta_0\right) - \Phi\left(\ell(u) - \beta_0\right) \\ \ell(u) &\equiv (-c - u)/\rho \\ r(u) &\equiv (c - u)/\rho \end{aligned}$$

Equal-tailed Selective Test

Conditional on A

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

Conditional on A^c

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

Constructing a Confidence Interval

Simply invert the test: find the values of β_0 that are not rejected.

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