Shrinkage regression for multivariate inference with missing data, and an application to portfolio balancing

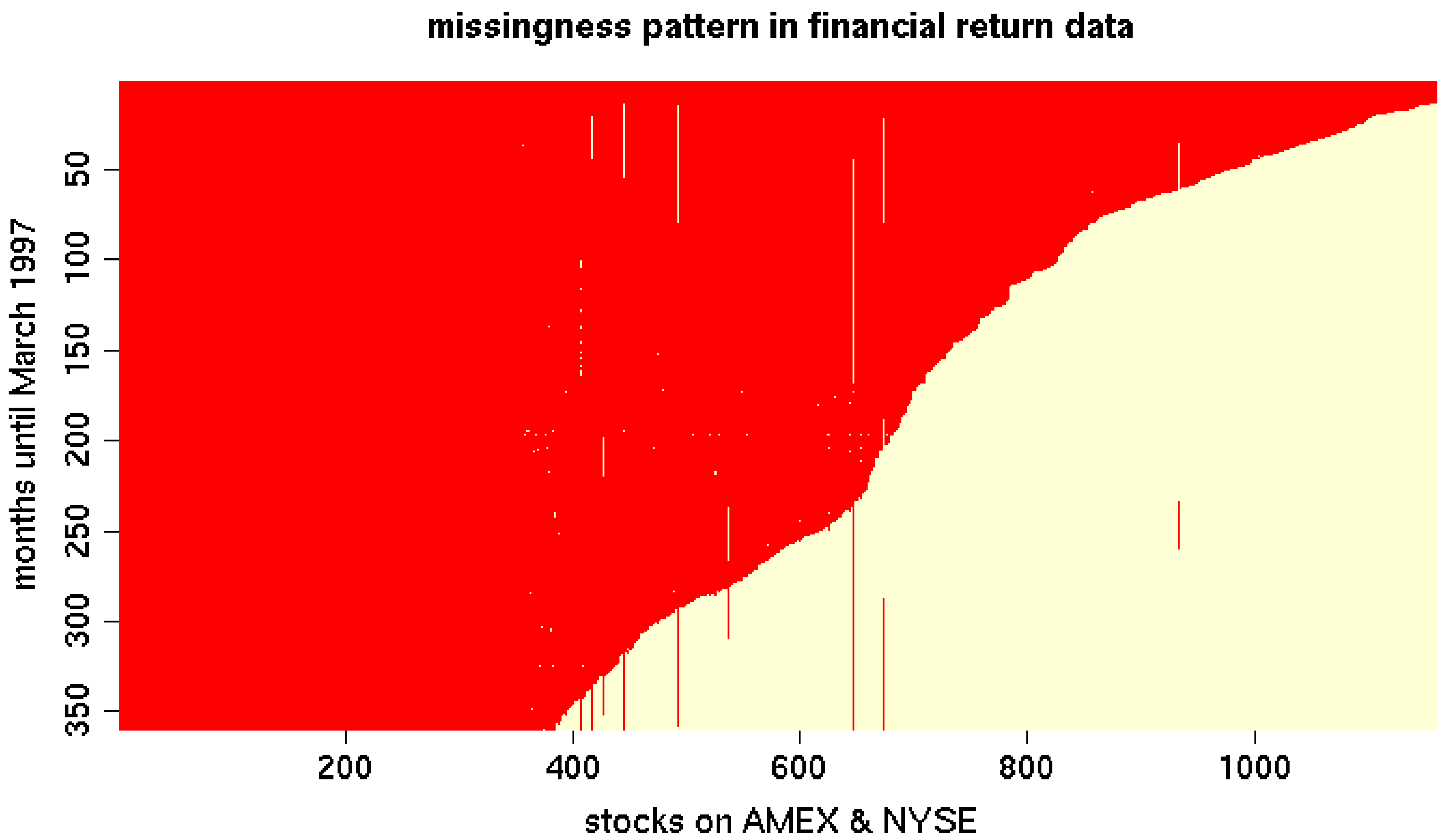
Robert B. Gramacy
Booth School of Business, The University of Chicago†
rbgramacy@chicagobooth.edu

Thanks to Joo Hee Lee, Ricardo Silva, and Ester Pantaleo

† Most of this work was done at the Statistical Laboratory, University of Cambridge

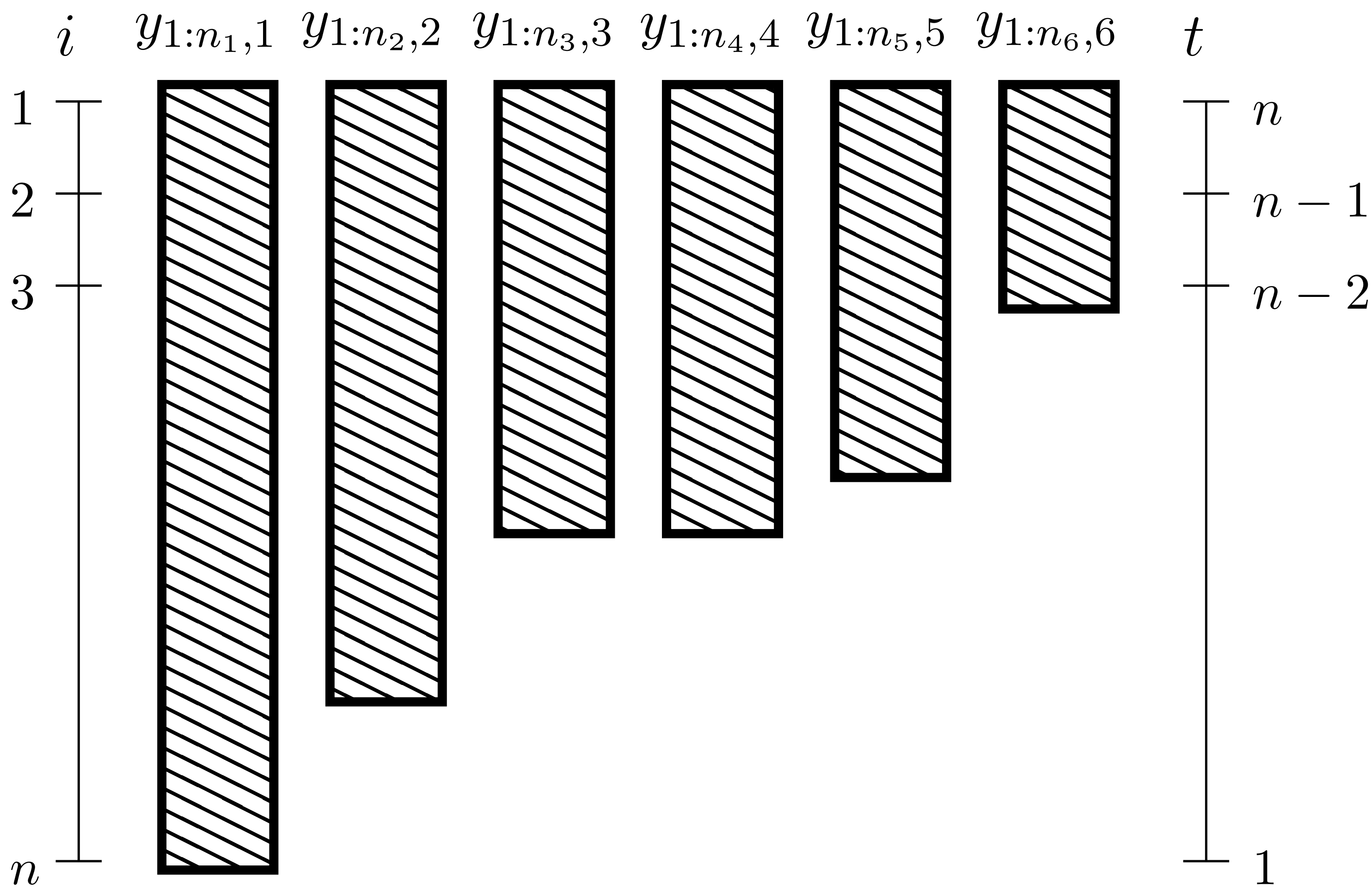
R **in Finance, UIC, April 2011**

NYSE & AMEX data from 1968–1997



- **Goal: to estimate MVN parameters** (μ, Σ)

Missingness pattern is *monotone*



$\mathbf{Y}$: $y_{:,1}, \ldots, y_{:,m}$ **and let** $\mathbf{y}_j \equiv y_{1:n_j,j}$

Easy to get MLE under MVN assumption

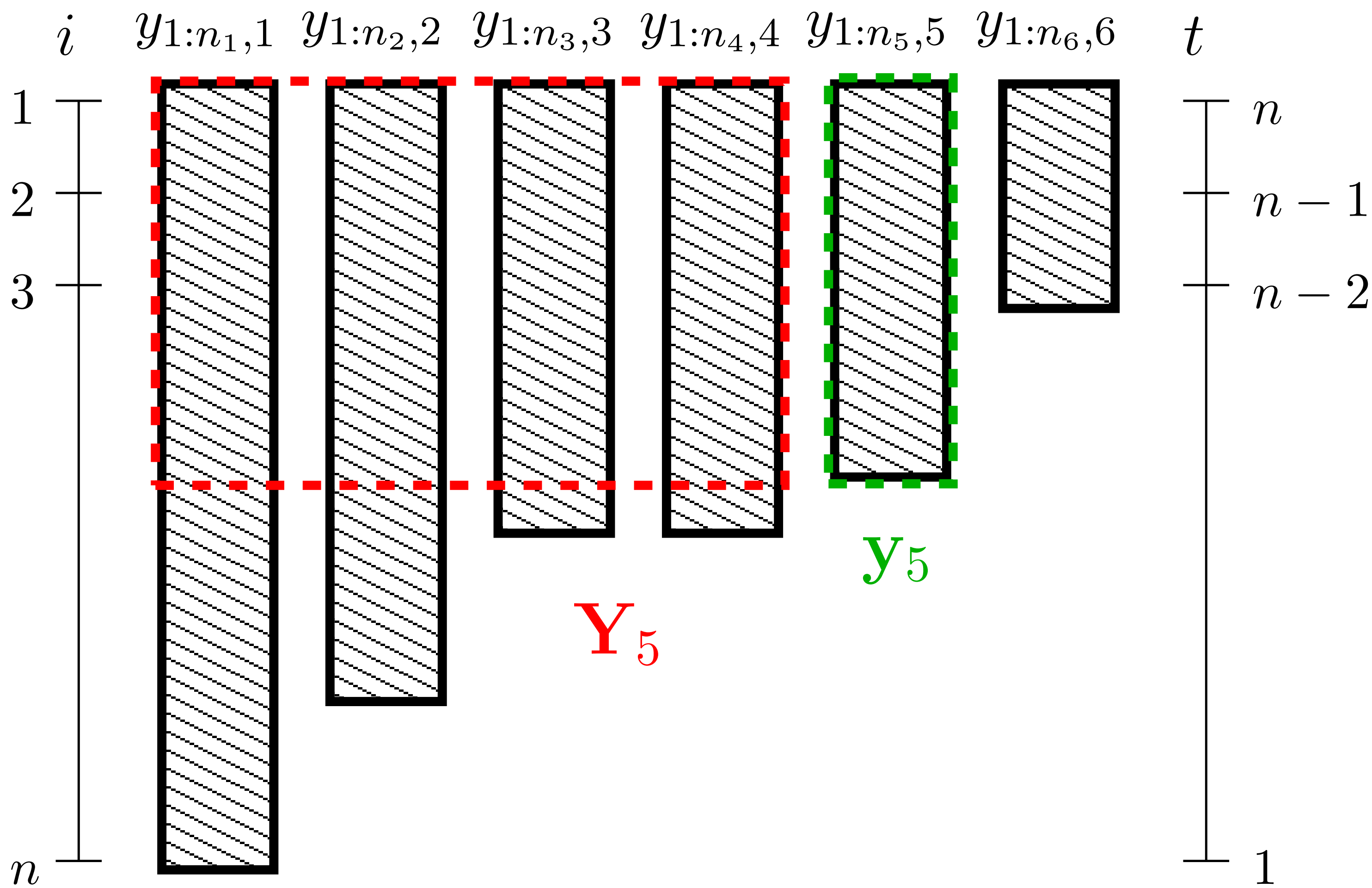
(Andersen 1957) **MLEs of $\boldsymbol{\theta}_j = (\mu_j, \boldsymbol{\Sigma}_{1:j,j})$, $j = 2, \dots, m$ may be obtained via OLS regressions**

$$\mathbf{y}_j = \mathbf{Y}_j \boldsymbol{\beta}_j + \boldsymbol{\epsilon}_j, \qquad \{\epsilon_{i,j}\}_{i=1}^{n_j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_j^2)$$

with $\phi_j = (\boldsymbol{\beta}_j, \sigma_j^2)$, where $\mathbf{y}_j \equiv y_{1:n_j,j}$ and

$$\mathbf{Y}_j \equiv \mathbf{Y}_{0:(j-1)}^{(n_j)} = \begin{pmatrix} 1 & y_{1,1} & \cdots & y_{1,(j-1)} \\ 1 & y_{2,1} & \cdots & y_{2,(j-1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & y_{n_j,1} & \cdots & y_{n_j,(j-1)} \end{pmatrix}$$

Repeated OLS regressions



$$\mathbf{y}_j = \mathbf{Y}_j \boldsymbol{\beta}_j + \boldsymbol{\epsilon}_j$$

MLE for OLS obtained in the usual way

- **When** $\mathrm{rank}(\mathbf{Y}_j) = j < n_j$**, OLS gives the MLE:**

$$\hat{\boldsymbol{\beta}}_j = (\mathbf{Y}_j^\top \mathbf{Y}_j)^{-1}\mathbf{Y}_j^\top \mathbf{y}_j \quad \textbf{and} \quad \hat{\sigma}_j^2 = \frac{1}{n_j}||\mathbf{y}_j - \mathbf{Y}_j\hat{\boldsymbol{\beta}}_j||^2$$

- $\hat{\boldsymbol{\theta}}_1 : \hat{\mu}_1 = \sum_{i=1}^{n_1} y_{i,1}/n_1$ **and** $\hat{\Sigma}_{1,1} = \sum_{i=1}^{n_1}(y_{i,1} - \hat{\mu}_1)^2/n_1$
- **Obtain** $\hat{\boldsymbol{\theta}}_j$ **from** $\hat{\boldsymbol{\theta}}_{1:(j-1)}$ **and** $\hat{\boldsymbol{\phi}}_j = (\hat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2)$ **as**

$$\hat{\mu}_j = \hat{\beta}_{0,j} + \hat{\boldsymbol{\beta}}_{1:(j-1),j}^\top \hat{\boldsymbol{\mu}}_{1:(j-1)}$$

$$\hat{\boldsymbol{\Sigma}}_{1:j,j} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_{1:(j-1),j}^\top \hat{\boldsymbol{\Sigma}}_{1:(j-1),1:(j-1)} \\ \hat{\sigma}_j^2 + \hat{\boldsymbol{\beta}}_{1:(j-1),j}^\top \hat{\boldsymbol{\Sigma}}_{1:(j-1),1:(j-1)} \hat{\boldsymbol{\beta}}_{1:(j-1),j} \end{pmatrix}$$

thus describing the mapping $\boldsymbol{\theta}_j = \Phi^{-1}(\boldsymbol{\theta}_{1:(j-1)}, \boldsymbol{\phi}_j)$

Example on cement data

Heat (y) evolved in setting of cement, as a function of its chemical composition ($x_{1:4}$) (Little & Rubin, 2002)

	original ordering				
n	x_1	x_2	x_3	x_4	y
1	7	26	6	60	78.50
2	1	29	15	52	74.30
3	11	56	8	20	104.30
4	11	31	8	47	87.60
5	7	52	6	33	95.90
6	11	55	9	22	109.20
7	3	71	17		102.70
8	1	31	22		72.50
9	2	54	18		93.10
10			4		115.90
11			23		83.80
12			9		113.30
13			8		109.40

$\Rightarrow$

	monotone ordering				
n	x_3	y	x_1	x_2	x_4
1	6	78.50	7	26	60
2	15	74.30	1	29	52
3	8	104.30	11	56	20
4	8	87.60	11	31	47
5	6	95.90	7	52	33
6	9	109.20	11	55	22
7	17	102.70	3	71	
8	22	72.50	1	31	
9	18	93.10	2	54	
10	4	115.90			
11	23	83.80			
12	9	113.30			
13	8	109.40			

The Bayesian approach

- **E.g., the popular non–informative prior**

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\Sigma|^{-\left(\frac{m+1}{2}\right)} \quad \Rightarrow \quad p(\boldsymbol{\beta}_j, \sigma_j^2) \propto (\sigma_j^2)^{-\left(\frac{m+1}{2} - m + j\right)}$$

for $j = 1, \ldots, m$, leads to the convenient posterior:

$$\boldsymbol{\beta}_j | \sigma_j^2, \boldsymbol{y}_j, \boldsymbol{Y}_j \sim \mathcal{N}_{j+1}(\hat{\boldsymbol{\beta}}_j, \sigma_j^2 (\mathbf{Y}_j^\top \boldsymbol{Y}_j)^{-1})$$

$$\sigma_j^2 | \mathbf{y}_j, \mathbf{Y}_j \sim \mathrm{IG}\left(\frac{n_j - m + j - 1}{2}, \frac{||\mathbf{y}_j - \mathbf{Y}_j \hat{\boldsymbol{\beta}}_j||^2}{2}\right)$$

- **Samples from m pairs of full conditionals converted to samples of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ via Φ^{-1}** (Polson & Tew, 2000)

Estimation Risk/Parameter uncertainty

(Zellner & Chetty, 1965; Klein & Bawa, 1976)

The posterior predictive distribution:

$$p(\mathbf{y}^{(t+1)}|\mathbf{Y}^{(t)}) = \int p(\mathbf{y}^{(t+1)}|\boldsymbol{\mu}, \boldsymbol{\Sigma})p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{Y}^{(t)})\, d\boldsymbol{\mu} d\boldsymbol{\Sigma}$$

moments $(\boldsymbol{\mu}^{(t+1)} = \hat{\mu}, \boldsymbol{\Sigma}^{(t+1)})$ **available w/o sampling**

- **no missing data:** $\boldsymbol{\Sigma}^{(t+1)} = c\hat{\boldsymbol{\Sigma}}$ (Polson & Tew, 2000)
- $\boldsymbol{\Sigma}^{(t+1)}$ **via** $\hat{\Sigma}$ **and** $\{n_j - j\}_{j=1}^m$ (Stambaugh, 1997)

Or, via samples from the posterior: (Polson & Tew, 2000)

$$\boldsymbol{\Sigma}^{(t+1)} = \mathbb{E}\{\boldsymbol{\Sigma}|\mathbf{Y}^{(t)}\} + \mathrm{Var}\{\boldsymbol{\mu}|\mathbf{Y}^{(t)}\}$$

The methods fail when

$\mathrm{rank}(\mathbf{Y}_j) = j \geq n_j$, **precluding** $(\mathbf{Y}_j^\top \mathbf{Y}_j)^{-1}$ **called a "big** p **small** n**" problem**

- **more parameters/predictors (**p**) :** $\texttt{ncol}(\mathbf{Y}_j) = j$
- **than observations (**n**) :** $\texttt{ncol}(\mathbf{Y}_j) = n_j$

Therefore for MLE/posterior, we cannot have:

- **an asset with fewer returns (**n_j**) than the number of assets with more returns (**$j-1$**)**
- **more assets than returns**

One solution: shrinkage regression

Instead of OLS we can obtain $\hat{\beta}$ and $\hat{\sigma}^2$ w/o $(\mathbf{X}^\top \mathbf{X})^{-1}$ via

$$\hat{\boldsymbol{\beta}}^{(q)} = \underset{\boldsymbol{\beta}}{\mathrm{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\}$$

where $\mathbf{y} \equiv \mathbf{y}_j$, $\mathbf{X} \equiv \mathbf{Y}_j$, and $\sigma^2 \equiv \sigma_j^2$.

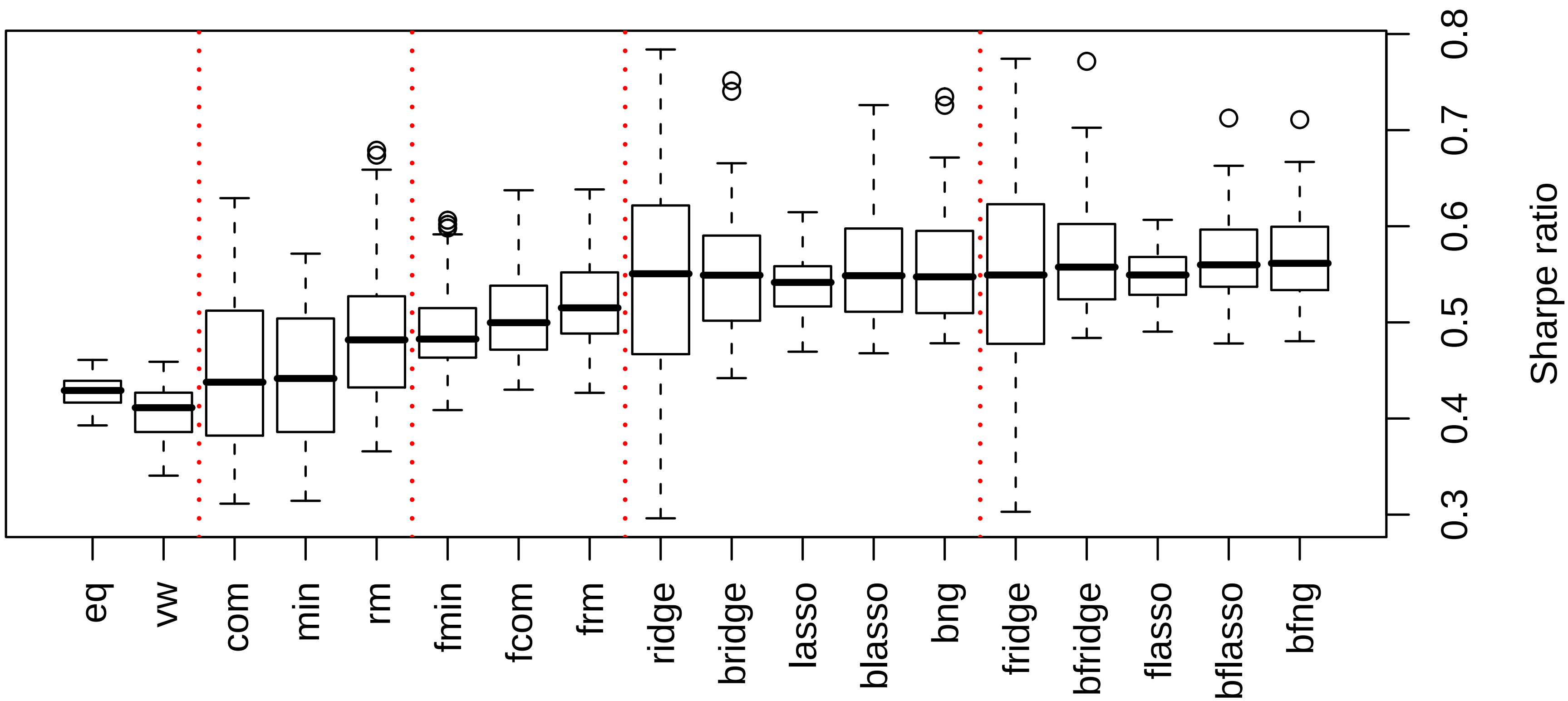
- $q = 2$ **(ridge);** $q = 1$ **(lasso)**

The shrinkage parameter, λ, may be chosen by CV

- **but we can can't account for estimation risk analytically** (Stambaugh, 1997) **via** $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = \Phi^{-1}(\hat{\phi})$
- **but with a fully Bayesian model we can sample!**

Monte Carlo investment exercise

- **Results of classical–Bayesian comparison**



Estimation Risk Matters

Failing to incorporate parameter uncertainty into the decision leads to lower quality investments

	Sharpe ratio	
Bayesian method	$\mathbb{E}\{\boldsymbol{\Sigma}\|\mathbf{Y}\}$	$\boldsymbol{\Sigma}^{(t+1)}$
Ridge	0.549	0.554
Ridge + Factor	0.562	0.571
Lasso	0.554	0.561
Lasso + Factor	0.562	0.573
NG	0.553	0.560
NG + Factor	0.563	0.574

Discussion and Implementation

- **extended** (Stambaugh, 1996) **to many assets**
- **even when OLS suffices, shrinkage has merits**
- **easy to relax MVN assumption via scale–mixtures**
- **easy to extend to the horseshoe**
- **even better for *mean*–variance portfolios**

`monomvn` **is made available as an** R **package**
Within R **do:**

```
R> install.packages(c("monomvn","lars","pls"))      #(once)
R> library(monomvn)
```