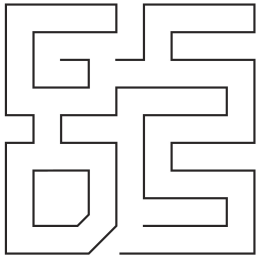
Graduate School
of Decision Sciences



Risky Oracles, Sparsity and Model Selection

4th Konstanz–Lancaster Workshop on Finance and Econometrics

Phillip Heiler

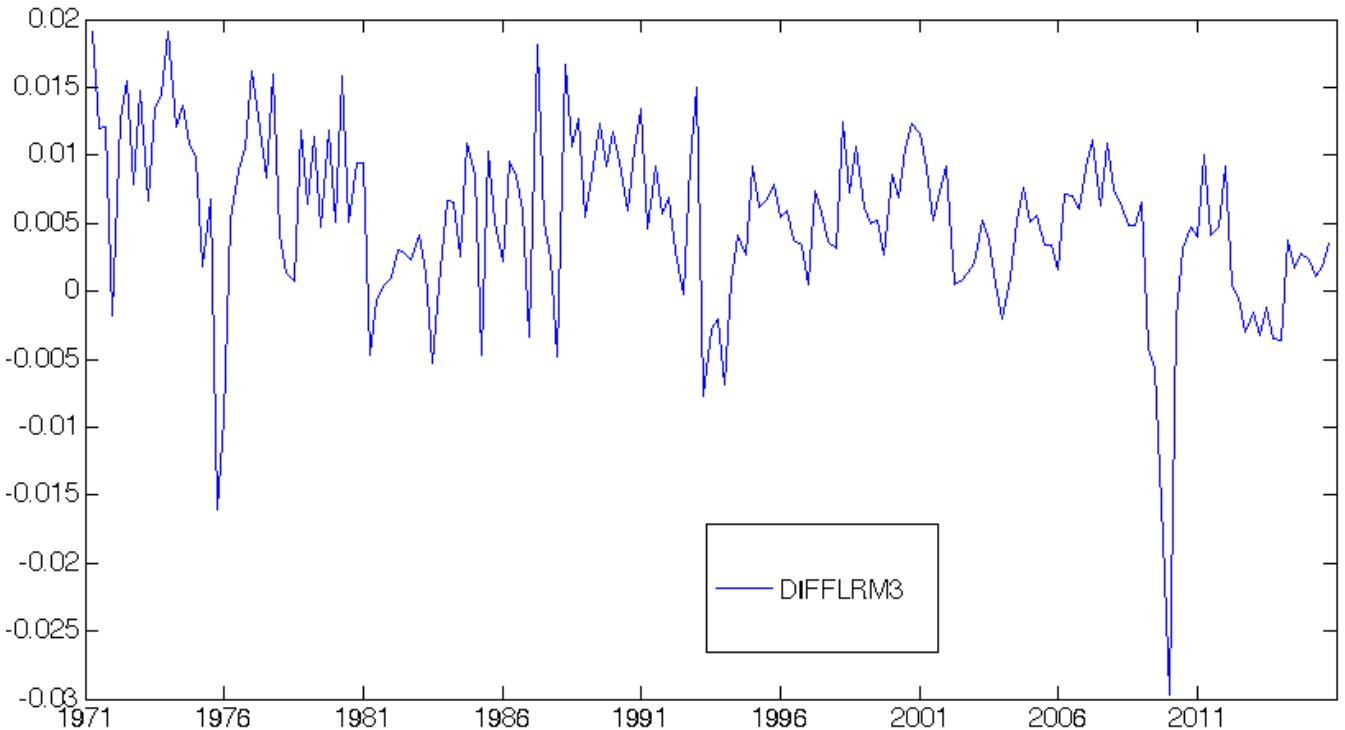
July 30, 2018

Table of Contents

A Typical Example (Applied Time-Series Analysis Exam)

Problem 5 (30 P) Consider quarterly data of first differences of log Euro Real M3 (deflated with the GDP deflator) for a period between 1970Q1 to 2014Q4 [...]:

Figure: Time series plot of $\Delta \log(RealM3_t)$



A Typical Example II (Applied Time-Series Analysis Exam)

d) **(2 P)** Table 2 shows results from applying information criteria for choosing the lag length of an autoregressive model for $\Delta \log(RealM3_t)$. What lag lengths are suggested by the information criteria? What lag would be your choice, why?

Lag	AIC (Akaike)	BIC (Schwarz)	HQ (Hannan-Quinn)
1	-7.471704	-7.436091	-7.378131
2	-7.479533	-7.426113	-7.339174
3	-7.469884	-7.398658	-7.282739
4	-7.458324	-7.369291	-7.224392
5	-7.460672	-7.353833	-7.179954

The Model Selection Problem: Literature

Model Selection: Box and Jenkins (1970), Mallow (1973), Aikaike (1973), Schwartz (1978), Shibata (1983), Li (1987), Tibshirani (1996), Shao (1997), Knight and Fu (2000), Fan and Li (2001), Zou (2006), Belloni et al. (2013, 2014), Chernozhukov et al. (2017)

Post-Model Selection Inference: Leeb and Pötscher (2005, 2006, 2008), Pötscher and Schneider (2009), Cattaneo et. al (2012)

Post-Model Selection Risk: Yang (2005), Leeb and Pötscher (2005, 2008), Hansen (2016)

The Canonical Model

Consider the linear regression model

$$Y_i = X_i'\theta + \varepsilon_i$$

with $\varepsilon_i \overset{i.i.d}{\sim} (0, \sigma^2)$ having a density with finite information, p fixed regressors with $n^{-1}\sum_{i=1}^{n} X_i X_i' \to Q$ positive definite with rank p.

- What is the "true model"?
- Which regressors to include?
- How to minimize parameter estimation risk?
- How to best predict Y_i?
- ...

One solution: "sparse model selection"

Sparse Model Selection

Definition: Sparsity-Type Estimation

Let $P_{n,\theta}$ denote the distribution of $Y_1, \ldots, Y_n$. For $\theta \in \mathbb{R}^p$ let $r(\theta)$ be $p \times 1$ with components $r_i(\theta) = 0$ if $\theta_i = 0$ and $r_i(\theta) = 1$ if $\theta \neq 0$. An estimator satisfies a sparsity-type condition if

$$P_{n,\theta}(r(\hat{\theta}) \leq r(\theta)) \to 1$$

for every $\theta \in \theta$ as $n \to \infty$.

Sparse Model Selection

Definition: Sparsity-Type Estimation

Let $P_{n,\theta}$ denote the distribution of $Y_1, \ldots, Y_n$. For $\theta \in \mathbb{R}^p$ let $r(\theta)$ be $p \times 1$ with components $r_i(\theta) = 0$ if $\theta_i = 0$ and $r_i(\theta) = 1$ if $\theta \neq 0$. An estimator satisfies a sparsity-type condition if

$$P_{n,\theta}(r(\hat{\theta}) \leq r(\theta)) \to 1$$

for every $\theta \in \theta$ as $n \to \infty$.

Examples:

- Complete subset selection/pre-testing,
- Bayesian information criteria,
- cross-validation with large validation set $n_v/n \to 1$,
- sparse estimation: SCAD, (adaptive) LASSO, ...

Sparse Model Selection II

Consistent model selection techniques fulfill

$$P_{n,\theta}(r(\hat{\theta}) = r(\theta)) \to 1$$

On first sight, sparsity seems like a desirable feature:

- Parameters are excluded if the true parameters are zero asymptotically.

However:

- Not informative about the actual finite-sample risk,
- risk will depend heavily on parameter values,
- can be very poor in finite samples.
- In fact: Maximal risk of **any** sparse estimator is maximally poor.

The Risk of Sparse Model Selection

For simplicity, consider the (scaled) MSE loss

$$l(\hat{\theta}, \theta) = n(\hat{\theta} - \theta)'(\hat{\theta} - \theta).$$

Recall that the risk of the ordinary least squares $\hat{\theta}_{OLS}$ is given by the expected loss

$$E_{n,\theta}[n(\hat{\theta}_{OLS} - \theta)'(\hat{\theta}_{OLS} - \theta)] = \sigma^2 tr\left(n^{-1}\sum_{i=1}^{n} X_i X_i'\right)$$
$$\rightarrow \sigma^2 tr(Q^{-1})$$

which remains bounded as $n \rightarrow \infty$ independently of the parameter values, i.e. also the maximum risk over the parameter space remains bounded. This is **not** the case for sparsity-type estimation.

The Risk of Sparse Model Selection II

Theorem 2.1. (Leeb and Pötscher, 2008)

Let $\hat{\theta}$ be an arbitrary estimator satisfying the sparsity-type condition. Then the maximal (scaled) mean squared error of $\hat{\theta}$ diverges to infinity, i.e.

$$\sup_{\theta \in \mathbb{R}^p} E_{n,\theta}[n(\hat{\theta} - \theta)'(\hat{\theta} - \theta)] \to \infty$$

as $n \to \infty$. More generally, let $l : \mathbb{R}^p \to \mathbb{R}$ be a nonnegative loss function. Then

$$\sup_{\theta \in \mathbb{R}^p} E_{n,\theta}[l(n^{1/2}(\hat{\theta} - \theta))] \to \sup_{s \in \mathbb{R}^p} l(s)$$

as $n \to \infty$.

The Risk of Sparse Model Selection III

Proof: Let $\theta_n = -n^{-1/2}s$, $s \in \mathbb{R}^p$ arbitrary, we have that

$$\begin{aligned}
\sup_{u \in \mathbb{R}^p} l(u) &\geq \sup_{u \in \mathbb{R}^p} E_{n,\theta} l(n^{1/2}(\hat{\theta} - \theta)) \\
&\geq E_{n,\theta_n} l(n^{1/2}(\hat{\theta} - \theta_n)) \\
&\geq E_{n,\theta_n}[l(n^{1/2}(\hat{\theta} - \theta_n))\mathbb{1}(\hat{\theta} = 0)] \\
&= l(-n^{1/2}\theta_n) P_{n,\theta_n}(r(\hat{\theta}) = 0) \\
&= l(s) P_{n,\theta_n}(r(\hat{\theta}) = 0).
\end{aligned}$$

Note that the sequence of probability measures P_{n,θ_n} is contiguous with $P_{n,0}$. By sparsity $P_{n,0}(r(\hat{\theta}) = 0) \to 1$. Since s is arbitrary, the proof is complete.

The Risk of Sparse Model Selection IV

$$\sup_{\theta \in \mathbb{R}^p} E_{n,\theta}[l(n^{1/2}(\hat{\theta} - \theta))] \to \sup_{s \in \mathbb{R}^p} l(s)$$

- Maximal risk of any sparse estimator is as bad as possible,
- remains true over open balls ρ_n centered at 0 as long as $n^{1/2}\rho_n \to \infty$
- bad risk behavior occurs exactly around the point where we would expect largest gain over e.g. OLS due to sparsity.
- also holds for data-dependent sparsity rule (Yang, 2005).

$\Rightarrow$ Fundamental conflict between sparse model selection and maximum risk.

A hypothetical conversation:

Leeb:

"Thus any sparsity-based estimator has the worst maximum risk possible."

A hypothetical conversation:

Leeb:
"Thus any sparsity-based estimator has the worst maximum risk possible."
Yang:
"...and this fundamental conflict cannot be resolved."

A hypothetical conversation:

Leeb:
"Thus any sparsity-based estimator has the worst maximum risk possible."
Yang:
"...and this fundamental conflict cannot be resolved."
Tibshirani:
"But what about the risk of SCAD and (adaptive) LASSO?"

A hypothetical conversation:

Leeb:
"Thus any sparsity-based estimator has the worst maximum risk possible."
Yang:
"...and this fundamental conflict cannot be resolved."
Tibshirani:
"But what about the risk of SCAD and (adaptive) LASSO?"
Fan:
"Yes, we showed that they reach near oracle performance."

A hypothetical conversation:

Leeb:
"Thus any sparsity-based estimator has the worst maximum risk possible."
Yang:
"...and this fundamental conflict cannot be resolved."
Tibshirani:
"But what about the risk of SCAD and (adaptive) LASSO?"
Fan:
"Yes, we showed that they reach near oracle performance."
B. Hansen:
"Yes, but your oracle performance also depends on the location of your parameters. It suffers from the same drawback."

(Near) Oracle Performance of the LASSO

Consider the orthonormal linear regression model

$$Y_i = X_i'\theta + \varepsilon_i$$

with $\varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$, p fixed regressors with $n^{-1}\sum_{i=1}^n X_i X_i' = I_p$. The LASSO is the solution to the following optimization problem:

$$\hat{\theta}_L = \arg\min_{\theta \in \mathbb{R}^p} \sum_{i=1}^n (Y_i - X_i'\theta)^2 + \lambda \sum_{j=1}^p |\theta_j|$$

and has a closed form solution in the orthonormal design, i.e.

$$\hat{\theta}_L = (t(\hat{\theta}_1), \ldots, t(\hat{\theta}_p))$$

with $t(x) = sign(x)(|x| - \lambda)_+$ ("soft-thresholding").

(Near) Oracle Performance of the LASSO II

Let $\sigma_n^2 = \sigma^2/n$. We are interested in the risk of the LASSO relative to OLS:

$$\rho(\hat{\theta}_L, \theta) = \frac{E[(\hat{\theta}_L - \theta)'(\hat{\theta}_L - \theta)]}{E[(\hat{\theta}_{OLS} - \theta)'(\hat{\theta}_{OLS} - \theta)]} = \frac{E[(\hat{\theta}_L - \theta)'(\hat{\theta}_L - \theta)]}{\sigma_n^2 p}$$

Donoho and Johnstone (1994) show that with optimal tuning $\lambda = \sigma_n\sqrt{2\ln p}$ the LASSO risk is bounded by

$$\rho(\hat{\theta}_L, \theta) \leq (1 + 2\ln p)\left(\frac{1}{p} + c_n(\theta)\right)$$

$$c_n(\theta) = \frac{1}{p}\sum_{j=1}^{p} \min\left\{\frac{\theta_j^2}{\sigma_n^2}, 1\right\}$$

Usual interpretation: Risk of LASSO close to oracle "kill-it-or-keep-it" estimator, i.e. OLS with regressors $\theta_j/\sigma_n^2 \geq 1$.

(Near) Oracle Performance of the LASSO III

$$\rho(\hat{\theta}_L, \theta) \leq (1 + 2 \ln p)\left(\frac{1}{p} + c_n(\theta)\right)$$

$$c_n(\theta) = \frac{1}{p} \sum_{j=1}^{p} \min\left\{\frac{\theta_j^2}{\sigma_n^2}, 1\right\}$$

Potential problem of the oracle interpretation:

- $\ln(p)$-term,
- **"kill-it-or-keep-it" estimator is neither oracle nor optimal**,
- its performance crucially depends on the parameter values,

(Near) Oracle Performance of the LASSO IV

Case 1: All coefficients of similar magnitude and large, i.e.

$$\theta_j^2/\sigma_n^2 > 1$$

for all j. The relative risk bound is then given by

$$\rho(\hat{\theta}_L, \theta) \leq (1 + 2\ln p)\left(\frac{1}{p} + \frac{1}{p}\sum_{j=1}^{p} \min\left\{\frac{\theta_j^2}{\sigma_n^2}, 1\right\}\right)$$

$$= (1 + 2\ln p)\left(\frac{1}{p} + 1\right)$$

which for large p is approximately

$$(1 + 2\ln p)$$

and thus the LASSO can be worse than OLS.

(Near) Oracle Performance of the LASSO V

Case 2: All coefficients of similar magnitude and small, i.e.

$$\theta_j^2/\sigma_n^2 = c \leq 1$$

for all j. The relative risk bound is then given by

$$\rho(\hat{\theta}_L, \theta) \leq (1 + 2\ln p)\left(\frac{1}{p} + \frac{1}{p}\sum_{j=1}^{p} \min\left\{\frac{\theta_j^2}{\sigma_n^2}, 1\right\}\right)$$

$$= (1 + 2\ln p)\left(\frac{1}{p} + c\right)$$

which for large p is approximately

$$(1 + 2\ln p)c$$

and thus the LASSO can be worse than OLS for some c.

(Near) Oracle Performance of the LASSO VI

Case 3 ("Sparse regression"): $k < p$ coefficients are large, the remaining 0, i.e.

$$\theta_j^2/\sigma_n^2 \begin{cases} \geq 1 & \text{for } j = 1, \ldots, k \\ = 0 & \text{for } j = k+1, \ldots, p. \end{cases}$$

The relative risk bound is then given by

$$\begin{aligned} \rho(\hat{\theta}_L, \theta) &\leq (1 + 2 \ln p)\left(\frac{1}{p} + \frac{1}{p}\sum_{j=1}^{p} \min\left\{\frac{\theta_j^2}{\sigma_n^2}, 1\right\}\right) \\ &= (1 + 2 \ln p)\left(\frac{1+k}{p}\right) \end{aligned}$$

which is small for a fixed k and p large and thus the LASSO can dominate OLS.

Simulations: LASSO Finite Sample Risk

Simulation Design:

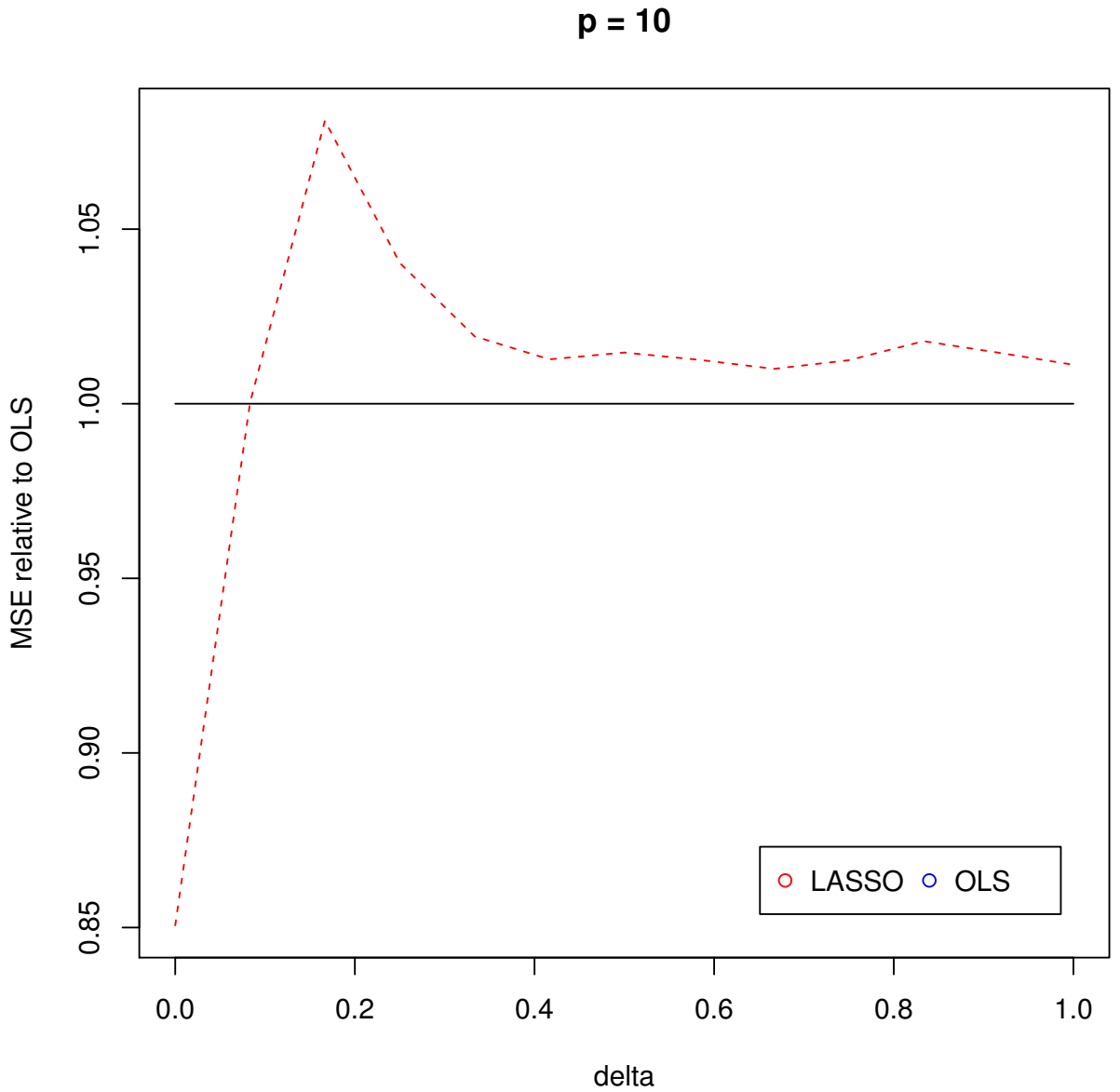
$$Y = X1 + X2\delta + \varepsilon$$

with

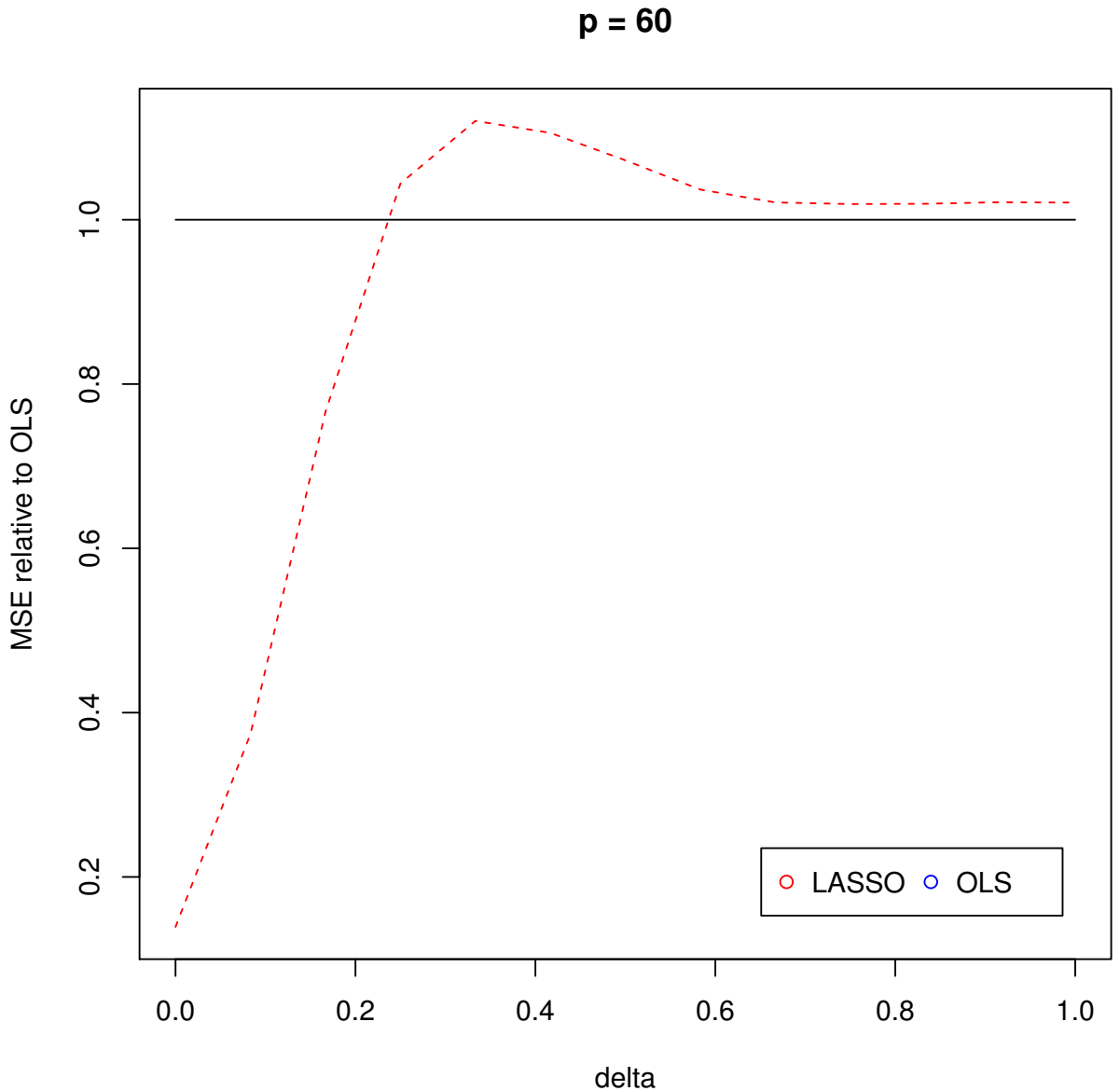
$$rk(X1) = 4$$
$$rk(X2) = p - 4$$

All entries of the columns of $X1$ and $X2$ and of ε are iid $\mathcal{N}(0, 1)$.
δ varies from 0 (sparse regression) to 1 (only big coefficients).
$n = 100$. We compare the risk of LASSO with 5-fold cross-validation to OLS.

Simulations: LASSO Finite Sample Risk II



Simulations: LASSO Finite Sample Risk III



Extension: Subvector Risk

Say instead we are only interested in the risk of subvector θ_a when $\theta = (\theta_a, \theta_b)$. Assume we perform sparse model selection of θ_b:

- worst maximum risk for θ_b
- worst maximum risk for θ_a for most classical model selection techniques

However minimax risk attainable for θ_a for some selection methods under sufficient sparsity for θ_b:

- double-selection (Belloni et al. 2013, 2014)
- double machine learning (Chernozhukov et al. 2017)

Conclusion

- Goal of the model selection step is relevant.
- unsolvable trade-off between correct model selection and estimation risk,
- for both classical and "modern" model selection,
- oracle asymptotic risk properties have to be read correctly,
- sparse model selection methods have undesirable maximum risk properties,
- require thorough investigation of the parameter space,
- possibility for minmax risk for unselected subvector components.

Thank you for your attention!