Econometric Reviews

Editor: Esfandiar Maasoumi

Volume 24 Number 3 2005

Taylor & Francis
Taylor & Francis Group

Econometric Reviews

ISSN: 0747-4938 (Print) 1532-4168 (Online) Journal homepage: http://tandfonline.com/loi/lecr20

# The Risk of James–Stein and Lasso Shrinkage

Bruce E. Hansen

**To cite this article:** Bruce E. Hansen (2016) The Risk of James–Stein and Lasso Shrinkage, Econometric Reviews, 35:8-10, 1456-1470, DOI: 10.1080/07474938.2015.1092799

**To link to this article:** http://dx.doi.org/10.1080/07474938.2015.1092799

Accepted author version posted online: 13 Oct 2015.
Published online: 13 Oct 2015.

*Econometric Reviews*, 35(8–10):1456–1470, 2016
Copyright © Taylor & Francis Group, LLC
ISSN: 0747-4938 print/1532-4168 online
DOI: 10.1080/07474938.2015.1092799

# The Risk of James–Stein and Lasso Shrinkage

Bruce E. Hansen

*Department of Economics, University of Wisconsin, Madison, Wisconsin, USA*

This article compares the mean-squared error (or $\ell_2$ risk) of ordinary least squares (OLS), James–Stein, and least absolute shrinkage and selection operator (Lasso) shrinkage estimators in simple linear regression where the number of regressors is smaller than the sample size. We compare and contrast the known risk bounds for these estimators, which shows that neither James–Stein nor Lasso uniformly dominates the other. We investigate the finite sample risk using a simple simulation experiment. We find that the risk of Lasso estimation is particularly sensitive to coefficient parameterization, and for a significant portion of the parameter space Lasso has higher mean-squared error than OLS. This investigation suggests that there are potential pitfalls arising with Lasso estimation, and simulation studies need to be more attentive to careful exploration of the parameter space.

**Keywords** James-Stein; Lasso; Least-squares; Shrinkage.

**JEL Classification** C13.

## 1. INTRODUCTION

Two important and quite distinct shrinkage estimators are the James–Stein estimator (James and Stein, 1961) and the least absolute shrinkage and selection operator (Lasso) (Tibshirani, 1996). The Lasso in particular has become quite popular in the recent literature, supported in part by oracle results concerning its ability to nearly achieve the risk (mean-squared-error) of infeasible optimal selection in canonical regression. This article explores and compares the finite sample risk of these two estimators.

The idea of James–Stein shrinkage dates back to the seminal article of Stein (1956) which showed that Gaussian estimators are inadmissible when the dimension exceeds two. James and Stein (1961) provided a constructive shrinkage estimator which dominates the conventional estimator, and Baranchick (1964) shows that positive part trimming further reduces the risk. The theoretical foundation for risk calculations was provided by Stein (1981) and is carefully explored in Chapter 5 of Lehmann and Casella (1998).

Address correspondence to Dr. Bruce E. Hansen, Department of Economics, University of Wisconsin, 1180 Observatory Drive, Madison, WI 53706, USA; E-mail: behansen@wisc.edu

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lecr.

The theory of efficient high-dimensional shrinkage was developed by Pinsker (1980), and the connection to the James–Stein estimator is succinctly explained in Chapter 7 of Wasserman (2006).

The Lasso (Tibshirani, 1996) minimizes the sum of squared errors subject to an $\ell_1$ penalty. Related estimators include the non-negative garotte (Breiman, 1995), Smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), elastic net (Zou and Hastie, 2005), adaptive Lasso (Zou, 2006), and ordinary least squares (OLS) post-Lasso (Belloni and Chernozhukov, 2013). In this article, we decided to focus on just two of these estimators: the Lasso, and OLS post- Lasso. This is partially so as there is a well-developed oracle inequality for the risk of these estimators in the canonical regression model (Donoho and Johnstone, 1994), partially because there is a widely-accepted method for selection of the penalty parameter (5-fold cross-validation), and partially as Belloni and Chernozhukov (2013) have made a strong case for OLS post-Lasso as a method to reduce the bias of Lasso and thus possibly its risk.

An important property of the Lasso is that it can be applied even when the number of regressors $p$ exceeds the number of observations $n$. This is not the case for OLS and James–Stein shrinkage, both of which require $p < n$. Since our goal is to compare these two shrinkage methods, our focus is exclusively on the case $p < n$.

This article points to some limitations of the Lasso method. Other articles that have also pointed out limitations concerning inference using Lasso estimation include Pötscher and Leeb (2008, 2009) and Pötscher and Schneider (2009). Another related work is Cattaneo et al. (2012), which explores inference in regression models with $p$ proportional to $n$.

This article does not contribute to the theory of shrinkage estimation. Instead, it reviews what is known concerning the finite sample risk of the shrinkage estimators, and then carefully explores the mean-squared error in a simple simulation experiment.

The organization of the article is as follows. Section 2 describes four key estimators (OLS, James–Stein shrinkage, Lasso, and OLS post-Lasso). Section 3 presents bounds on the finite sample risk of the shrinkage estimators in the canonical regression model. Section 4 is the main contribution: a simulation study on the finite sample risk.

The R code which creates the simulation results is available on the author's website[1].

## 2. SHRINKAGE ESTIMATORS

Consider the linear regression

$$y_i = x_{0i}'\beta_0 + \sum_{j=1}^{p} x_{ji}\beta_j + e_i, \tag{1}$$

[1] www.ssc.wisc.edu/~bhansen/

$i = 1, \ldots, n$, where $x_{0i}$ is $k_0 \times 1$ and $x_{ji}$ is scalar for $j \geq 1$. Assume $p \geq 3$ and $p + k_0 \leq n$, so that all estimators described below exist. Define $\beta = (\beta_0', \beta_1, \ldots, \beta_p)'$, $x_i = (x_{0i}', x_{1i}, \ldots, x_{pi})'$, and define the matrices $X_0$, $X$, and $Y$ by stacking observations.

The OLS estimate of $\beta$ is $\hat{\beta} = (X'X)^{-1} X'Y$.

Now consider estimates of $\beta$ which treat the coefficients $\beta_1, \ldots, \beta_p$ as possibly small. One choice is the constrained least-squares estimate subject to the constraint $\beta_1 = \beta_2 = \cdots = 0$, which is

$$\tilde{\beta} = \begin{pmatrix} (X_0'X_0)^{-1} X_0'Y \\ 0 \end{pmatrix}.$$

Alternatively, consider a shrinkage estimator which shrinks the unconstrained estimator $\hat{\beta}$ towards $\tilde{\beta}$. The (positive-part) James–Stein estimator is

$$\hat{\beta}^{JS} = \tilde{\beta} + \left(\hat{\beta} - \tilde{\beta}\right) \left(1 - \frac{(p-2)\,\hat{\sigma}^2}{\left(\hat{\beta} - \tilde{\beta}\right)' X'X \left(\hat{\beta} - \tilde{\beta}\right)}\right)_+$$

where $(a)_+ = \max[a, 0]$ is the positive-part operator and $\hat{\sigma}^2 = (n - k_0 - p)^{-1} \sum_{i=1}^n \left(y_i - x_i'\hat{\beta}\right)^2$ is the standard estimate of $\sigma^2 = Ee_i^2$. The James–Stein estimator $\hat{\beta}^{JS}$ is a weighted average of the unconstrained and constrained least-squares estimators, with the weight on the unconstrained estimator an increasing function of the distance measure $D_n = \left(\hat{\beta} - \tilde{\beta}\right)' X'X \left(\hat{\beta} - \tilde{\beta}\right) / \hat{\sigma}^2$. The James–Stein estimator can be viewed as a smoothed version of a pre-test estimator which selects the unrestricted estimator $\hat{\beta}$ when the statistic $D_n$ is large and the restricted estimator $\tilde{\beta}$ otherwise.

The Lasso estimator which shrinks $\hat{\beta}$ towards $\widetilde{\beta}$ solves

$$\hat{\beta}^L = \underset{\beta}{argmin} \frac{1}{2n} \sum_{i=1}^n \left(y_i - x_i'\beta\right)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

for some $\lambda > 0$. In practice, the estimator requires selection of the penalty parameter $\lambda$. A common choice is to pick $\lambda$ to minimize 5-fold cross-validation.

In the definitions $\hat{\beta}^{JS}$ and $\hat{\beta}^L$ the coefficients on the regressors $x_{0i}$ are not shrunk (and thus these regressors are always retained. Typical applications set $x_{0i} = 1$ (including the default setting in the R package **glmnet**), but any choice is feasible. For example, if $x_{0i}$ is set to null, then $\tilde{\beta}$ is the zero vector and the Lasso $\hat{\beta}^L$ shrinks all coefficients symmetrically, including the intercept.

One feature of the Lasso estimator $\hat{\beta}^L$ is that it simultaneously performs selection and shrinkage. That is, some of the individual coefficient estimates may equal zero, so that $\hat{\beta}_j^L = 0$ is possible for some $j$. Let $\widehat{S}$ be a selector matrix which selects the coefficients not set to zero, so that $X\widehat{S}$ are the regressors "selected" by the Lasso. The OLS post-Lasso estimator of Belloni and Chernozhukov (2013) is least-squares performed on these variables and can be written as

$$\hat{\beta}^P = \left(\widehat{S}'X'X\widehat{S}\right)^{-1}\widehat{S}'X'Y.$$

## 3. CANONICAL MODEL

Let us study the simplified setting of full shrinkage, orthogonal regressors, and normal errors with known variance. That is, $k_0 = 0$ (no $x_{0i}$), $e_i \sim N(0, \sigma^2)$, and $n^{-1}X'X = I_p$. In this setting, the James–Stein estimator equals

$$\hat{\beta}^{JS} = \hat{\beta}\left(1 - \frac{(p-2)\,\sigma_n^2}{\left(\hat{\beta} - \tilde{\beta}\right)'\left(\hat{\beta} - \tilde{\beta}\right)}\right)_+$$

where $\sigma_n^2 = \sigma^2/n$. (Notice that in this context the estimator is a function of the known variance $\sigma^2$ rather than the estimator $\hat{\sigma}^2$.)

The Lasso estimator equals

$$\hat{\beta}^L = \left(t(\hat{\beta}_1), t(\hat{\beta}_2), \ldots, t(\hat{\beta}_p)\right)',$$

where $t(x) = sign(x)\,(|x| - \lambda)_+$, and is also known as soft-thresholding estimator. See Tibshirani (1996, Eq. (3)), Fan and Li (2001, Eq. (2.6)), and van der Geer and Bühlmann (2011, p. 10). Note that orthogonality is essential for this simplification.

Similarly, the OLS post-Lasso estimator equals

$$\hat{\beta}^P = \left(s(\hat{\beta}_1), s(\hat{\beta}_2), \ldots, s(\hat{\beta}_p)\right)',$$

where $s(x) = x1\,(|x| > \lambda)$, and is also called the hard-thresholding estimator.

For any estimator $\bar{\beta}$ of $\beta$, we define the (normalized) mean-squared error or $\ell_2$ risk as

$$R\left(\bar{\beta}, \beta\right) = \frac{E\left(\bar{\beta} - \beta\right)'\left(\bar{\beta} - \beta\right)}{p\sigma_n^2}. \tag{2}$$

We have normalized by $p\sigma_n^2$ so that the risk of the OLS estimator is unity. Indeed,

$$\begin{aligned} R\left(\hat{\beta}, \beta\right) &= \frac{E\left((e'X)\,(X'X)^{-1}\,(X'X)^{-1}\,(X'e)\right)}{p\sigma_n^2} \\ &= \frac{E\,tr\,((X'e)\,(e'X))}{np\sigma^2} \end{aligned}$$

$$
\begin{aligned}
&= \frac{tr\left(I_p\right)}{p} \\
&= 1.
\end{aligned}
$$

Given the simplified setting, there are well-known bounds for the risk of the shrinkage estimators described above.

Take the James–Stein estimator. Its risk satisfies the bound

$$R\left(\hat{\beta}^{JS}, \beta\right) \leq \frac{2}{p} + \frac{c_n^{JS}(\beta)}{1 + c_n(\beta)}, \tag{3}$$

where

$$c_n^{JS}(\beta) = \frac{1}{p}\sum_{j=1}^{p}\frac{\beta_j^2}{\sigma_n^2}.$$

(See, for example, Wasserman, 2006, Theorem 7.42.)

When $p$ is large, we can see that the risk (3) is effectively bounded by $c_n^{JS}(\beta)/(1 + c_n^{JS}(\beta))$, which is strictly smaller than the risk of the least-squares estimator. The gains are highest (the relative risk smallest) when $c_n^{JS}(\beta)$ is small; equivalently when the normalized coefficients $\beta_1/\sigma_n, \ldots, \beta_p/\sigma_n$ are small.

Since $c_n^{JS}(\beta)/(1 + c_n^{JS}(\beta))$ is the Pinkser minimax bound for $\ell_2$ risk, the bound (3) shows that the James–Stein estimator is near minimax optimal when $p$ is large, as discussed in Chapter 7 of Wasserman (2006). It is optimal in the sense that its worst-case risk over any ball in $\beta$ centered at the the origin equals the minimax bound over the same region.

Next, take the Lasso estimator. For $\lambda = \sigma_n\sqrt{2\ln p}$, its risk satisfies the bound

$$R\left(\hat{\beta}^{L}, \beta\right) \leq (1 + 2\ln p)\left(\frac{1}{p} + c_n^L(\beta)\right), \tag{4}$$

where

$$c_n^L(\beta) = \frac{1}{p}\sum_{j=1}^{p}\min\left(\frac{\beta_j^2}{\sigma_n^2}, 1\right).$$

(Donoho and Johnstone, 1994, Theorem 2; see also Wasserman, 2006, Theorem 9.34.)

The OLS post-Lasso estimator also satisfies (4) for $\lambda$ close to $\sigma_n\sqrt{2\ln p}$, as established by Donoho and Johnstone (1994, Theorem 4).

The bound (4) is typically interpreted as showing that the risk of the Lasso estimator is close to $c_n^L(\beta)$. The latter is the risk of the infeasible "kill-it-or-keep-it" estimator which performs least-squares on the regressors whose coefficients satisfy $\beta_j^2/\sigma_n^2 \geq 1$. Thus the bound (4) has been interpreted as an oracle result, as it shows that the risk of the Lasso estimator is close to the risk of this infeasible estimator. This interpretation is a bit strained, however, not just because of the obvious $2 \ln p$ factor, but more importantly because the infeasible "kill-it-or-keep-it" estimator is itself neither oracle nor optimal. For example, the feasible James–Stein estimator has smaller risk than $c_n^L(\beta)$ when $\beta_1^2 = \cdots = \beta_p^2$, and thus $c_n^L(\beta)$ is not optimal, nor an efficiency bound.

A comparison of the bounds (3) and (4) shows that neither one uniformly dominates the other. Thus there are regions of the parameter space where the James–Stein estimator has lower risk, and regions of the parameter space where the Lasso estimator has lower risk. We can use these bounds to help us understand the regions where one estimator or the other has lower risk.

We can observe that $c_n^L(\beta) \leq c_n^{JS}(\beta)$, with equality only in the special case that $\beta_j^2 \leq \sigma_n^2$ for all $j$ (the setting that all coefficients are small). Suppose that this special case holds; then (3) is smaller than (4). Next, consider the other extreme, where $\beta_j^2 \geq \sigma_n^2$ for all $j$ (the setting that all coefficients are large). In this case, $c_n^L(\beta) = 1$, and again, we can see that (3) is smaller than (4). So, to roughly generalize, if all the coefficients are small in magnitude or all the coefficients are large, then the James–Stein estimator will have smaller risk than the Lasso estimator.

The case where (4) is smaller is when some coefficients are nonzero and some are zero (a *sparse* regression). In particular, suppose that $k < p$ coefficients satisfy $\beta_j^2 \geq \sigma_n^2$ and the remaining equal zero. Then $c_n^L(\beta) = k/p$ so the bound in (4) is

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

which approaches zero as $p \to \infty$ with $k$ fixed. In contrast, (3) can be arbitrarily close to one if the nonzero coefficients are large. In this context, (4) can be much smaller than (3).

The theory shows that there is no reason to expect either the James–Stein estimator nor the Lasso estimator to strictly dominate one another. The James–Stein estimator will have lower risk when the coefficients are roughly comparable in magnitude, while the Lasso estimator will have lower risk when a few coefficients are large in magnitude and the remainder are quite small. In mixed cases, it is more difficult to make an *a priori* ranking. Furthermore, these calculations are all for the case of orthogonal regressors, which is not generically relevant for empirical practice. To obtain a better picture in the next section, we turn to simulation estimates of the finite-sample risk.

## 4. SIMULATION EVIDENCE

### 4.1. Canonical Model

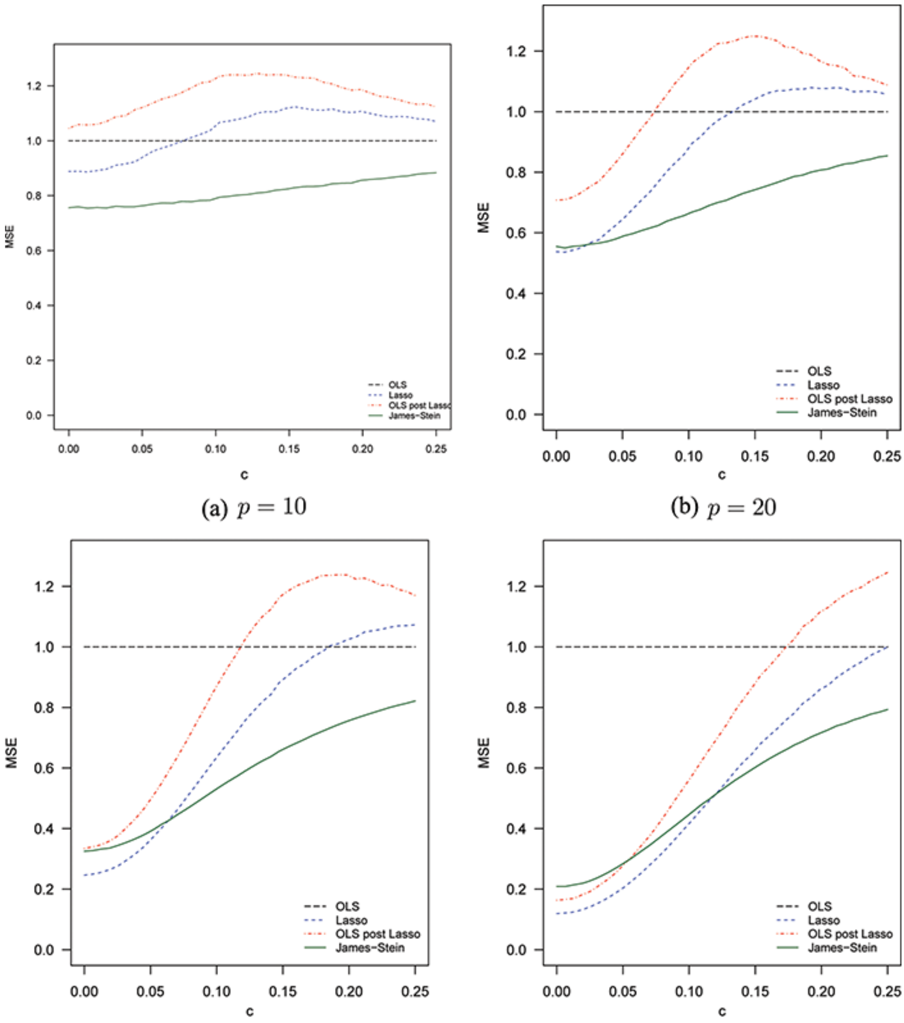
We calculate and compare the exact risk of estimates of model (1) in a finite sample simulation. We set $x_{0i} = 1$, the remaining regressors are mutually independent $N(0, 1)$, and the error $e_i$ is also independently distributed $N(0,1)$. We set $\beta_0 = 0$ and $n = 100$. Four "large" coefficients $\beta_j$ are set equal to a common value $b$. The remaining $p-4$ "small" coefficients $\beta_j$ are set equal to a common value $c$ which ranges from 0 to $b$. The value of $b$ is selected to target the population $R^2$ when $c = 0$, with $R^2$ varied between 0.05, 0.20 and 0.80. We view $R^2 = 0.20$ as a typical signal/noise ratio. The number of coefficients is varied between $p = 10, 20, 40$, and 60. The case $c = 0$ corresponds to a pure "sparse" design, while the case $c = b$ is the opposite extreme of equality of all coefficients. The intermediate case $0 < c < b$ is in between sparse and non-sparse designs, and is the interesting case where the relative rankings of James–Stein and Lasso estimation is *a priori* unclear.

To calculate the finite sample risk (2) of the estimators, we perform a simulation in R. We use 10,000 simulation replications and a 40-point grid for $c$. We compare the mean squared error (MSE) of four estimators: OLS, Lasso, OLS post-Lasso, and James–Stein. To implement the Lasso, we used the popular R package **glmnet** using all default settings, including orthogonalization of the regressors and nonshrinkage of the intercept. To select $\lambda$, we used 5-fold cross-validation as implemented in the companion package **cv.glmnet**. For the OLS post-Lasso estimator, we separately selected $\lambda$ by 5-fold cross-validation. For each estimator, we calculated the MSE, scaled by the MSE of the OLS estimator, so that values of relative MSE less than one indicate performance superior to OLS, and values above one indicating inferior performance to OLS.

Figure 1 displays the results for the case $R^2 = 0.20$. Each panel is for a distinct value of $p$, and the relative MSE graphed as a function of $c$.

Figure 1 shows that for $p = 10$ and $p = 20$, Lasso has considerably *higher* MSE than the simple James–Stein estimator, and that for many values of $c$, Lasso has even *higher* MSE than OLS. For larger values of $p$, Lasso has smaller MSE than James–Stein for small values of $c$, but the reverse holds for large values of $c$. Furthermore, Fig. 1 shows that OLS post-Lasso estimator does not generally have lower MSE than the Lasso estimator.

What Fig. 1 shows is that the enthusiastic embrace of Lasso estimation needs to be moderated. The methods works well in some contexts, but its accuracy is highly sensitive to the parameterization. The theory discussed in the previous section suggests that the Lasso is expected to work well in sparse settings (when many parameters are zero). However, we can see in Fig. 1 that for mild departures from spareness (when $c$ is small but positive) that Lasso can have higher MSE than the OLS estimator. The relative performance of Lasso estimation improves considerably as $p$ increases. For example, when $p = 60$ the James–Stein and Lasso estimators have similar MSE performance, but the MSE of Lasso estimation is more sensitive to the value of $c$.



(a) $p = 10$ (b) $p = 20$

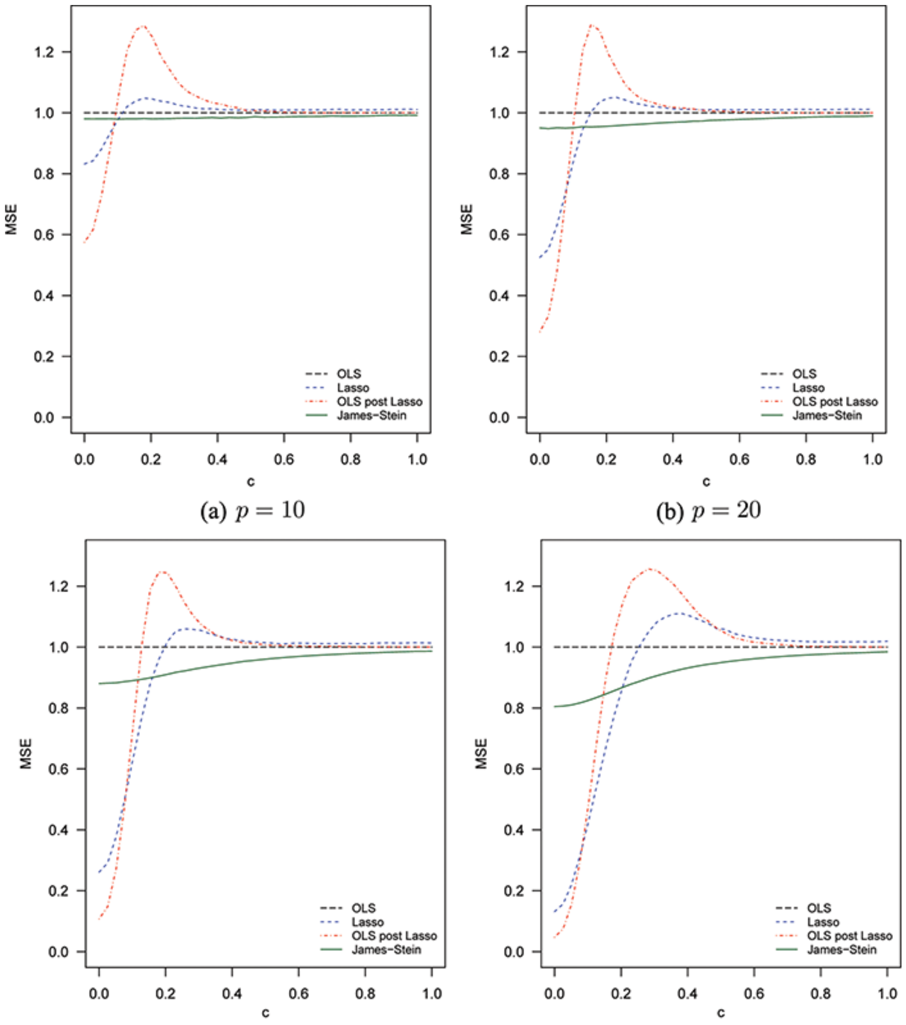
(c) $p = 40$ (d) $p = 60$

FIGURE 1 Canonical case, $R^2 = 0.2$.

## 4.2. Strong and Weak Signals

We next explore the consequence of the stronger signal-noise ratio $R^2 = 0.8$. Otherwise, the simulation experiment is the same as previously. We report the results in Fig. 2.

The comparisons are numerically different but qualitatively similar. Again, we find that the MSE of the Lasso estimator is highly sensitive to the value of $c$. It has low MSE when the design is sparse ($c$ is small), but for most values of $c$ has higher MSE than the



(a) $p = 10$

(b) $p = 20$

(c) $p = 40$

(d) $p = 60$

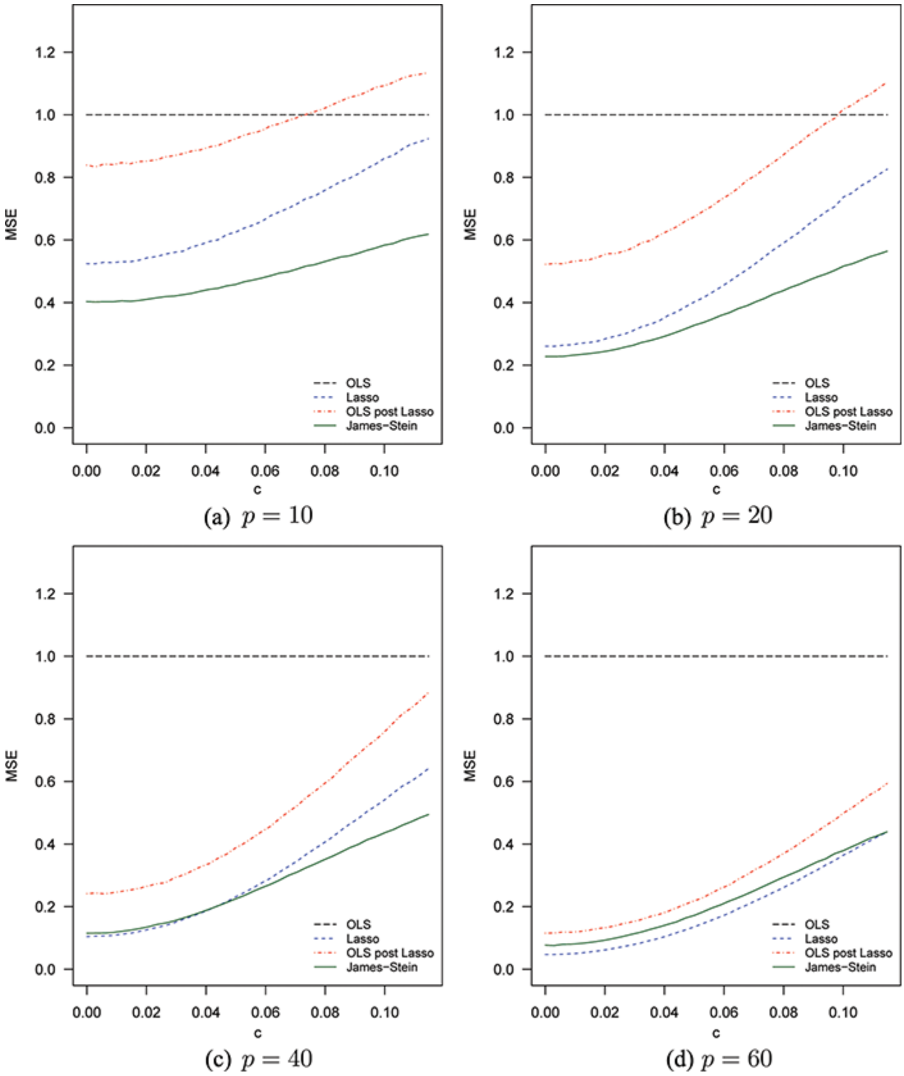
FIGURE 2 Canonical case, $R^2 = 0.8$.

OLS estimator. The James–Stein estimator uniformly dominates OLS, but only achieves a modest reduction in MSE.

Furthermore, in this setting the MSE of OLS post-Lasso is quite sensitive to the value of $c$. At $c = 0$, it has the smallest MSE (for all values of $p$), but for moderate values of $c$ its MSE is considerably higher than the others. This is similar to the MSE performance of classical pre-test estimators.

We next explore the consequence of the weak signal-noise ratio $R^2 = 0.05$. Otherwise, the simulation experiment is the same as previously. We report the results in Fig. 3. The results are qualitatively similar with Fig. 1.

Together, Figs. 1–3 show that the MSE of Lasso estimation is quite sensitive to the values of the parameters. Since it is *a priori* unknown if the true parameters are sparse or not, this is an undesirable property.



(a) $p = 10$

(b) $p = 20$

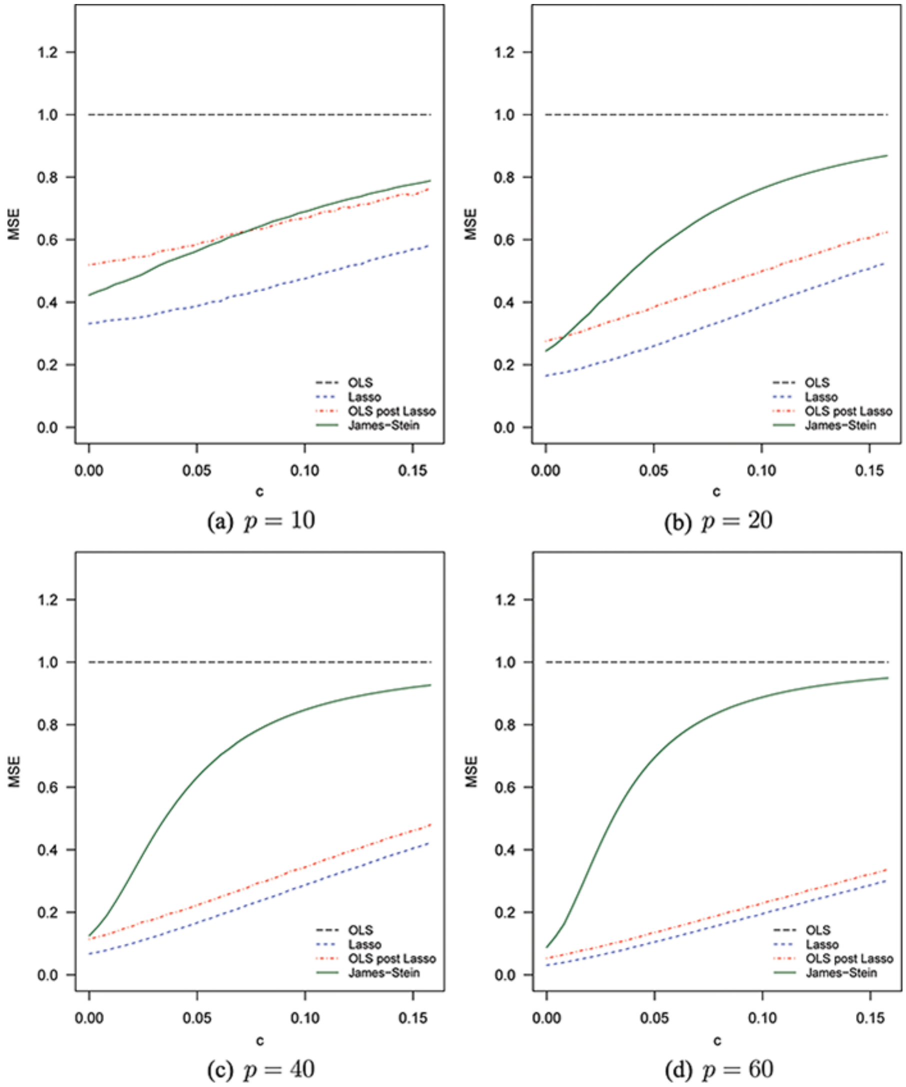
(c) $p = 40$

(d) $p = 60$

FIGURE 3 Canonical regression, $R^2 = 0.05$.

### 4.3. Correlated Regressors

We next explore the consequence of regressor correlation. For this experiment, the random variables in the regressor vector satisfy $x_i \sim N(0, \Sigma)$ with $\Sigma_{jj} = 1$ and $\Sigma_{jk} = 1/2$ for $j \neq k$. This is known as an "equi-correlated" regressor design. Setting $b$ so that $R^2 = 0.20$, the results are displayed in Fig. 4.



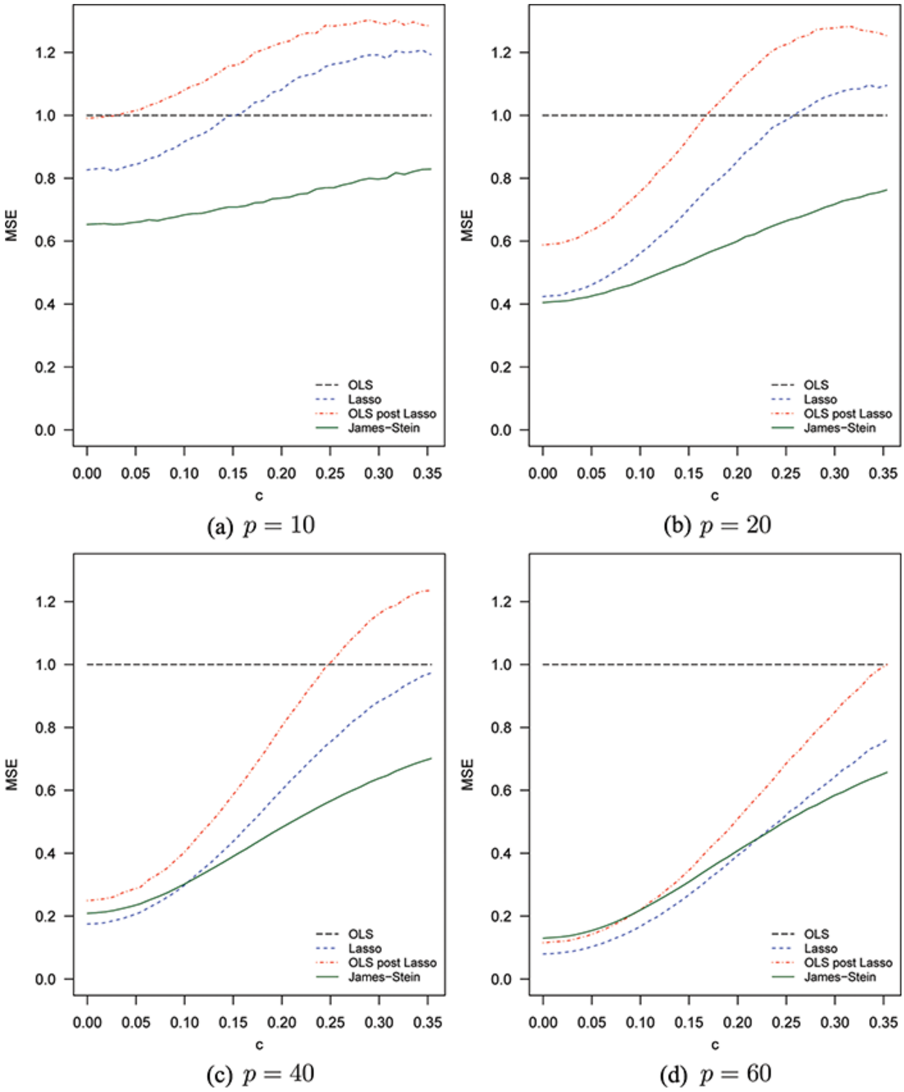
(a) $p = 10$

(b) $p = 20$

(c) $p = 40$

(d) $p = 60$

FIGURE 4 Equi-correlated regressors, $R^2 = 0.2$.



(a) $p = 10$

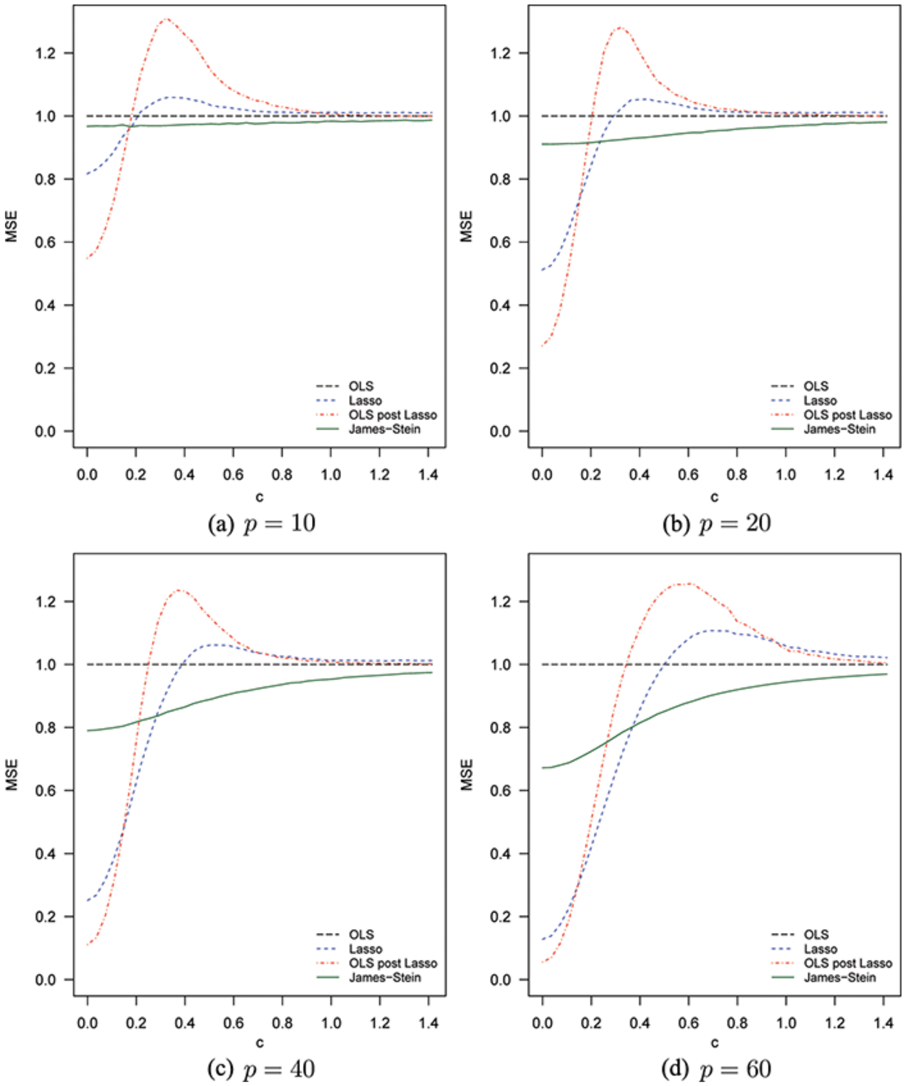
(b) $p = 20$

(c) $p = 40$

(d) $p = 60$

FIGURE 5 Equi-correlated regressors with alternating coefficient signs, $R^2 = 0.2$.

We find a reversal. In this experiment, the Lasso estimator has the smallest MSE, followed by OLS post-Lasso and then James–Stein. All estimators have considerably lower MSE than OLS.



(a) $p = 10$

(b) $p = 20$

(c) $p = 40$

(d) $p = 60$

FIGURE 6 Equi-correlated regressors, alternating coefficient signs, $R^2 = 0.8$.

It may be tempting to infer that Lasso has low MSE when the regressors are highly correlated. To show that this inference is fragile, we alter the coefficients so that one-half of them are positive and the other half negative. Thus we set

$$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \vdots \\ \beta_{p-1} \\ \beta_p \end{bmatrix} = \begin{bmatrix} b \\ -b \\ b \\ -b \\ c \\ -c \\ \vdots \\ c \\ -c \end{bmatrix}. \tag{5}$$

Otherwise, the experiment is the same as in Fig. 4. The results are presented in Fig. 5.

The rankings are reversed relative to Fig. 4, and are qualitatitively similar to the results from Fig. 1. For most parameter values, the James–Stein estimator has the smallest MSE, though for $p = 60$ the James–Stein and Lasso estimators have similar MSE.

Next, we explore the consequence of increasing the signal to $R^2 = 0.80$ in this setting. We report the results in Fig. 6.

The plots in Fig. 6 are similar to those of Fig. 2. Lasso has low MSE for small values of $c$, but has MSE higher than OLS for larger values of $c$. For most parameter values, the James–Stein estimator has the smallest MSE.

The results for the correlated regressor model shown in Figs. 4–6 complement our results for the canonical model. The relative performance of the Lasso estimator is quite sensitive to coefficient parameterization. For some parameterizations, it has low MSE, but for other parameterizations, it has higher MSE than OLS.

## 5. SUMMARY

The recent literature has devoted considerable attention and enthusiasm to Lasso estimation. It has many desirable properties, including the ability to handle cases where the number of regressors greatly exceeds the sample size. But what much of this literature has missed is that the performance of Lasso estimation depends critically upon the values of the coefficients. When the true coefficients satisfy a strong sparsity condition (there are many coefficients truly equal to zero), then Lasso estimation can have low MSE. But in other contexts the estimator need not have low MSE and can actually perform worse than simple OLS. Amid the hype, caution and skepticism appear warranted.

## ACKNOWLEDGEMENTS

I gratefully thank Jen-Che Liao for excellent research assistance, and two referees for helpful and constructive comments on an earlier draft.

## FUNDING

Research supported by the National Science Foundation.

## REFERENCES

Baranchick, A. (1964). Multiple regression and estimation of the mean of a multivariate normal distribution. Technical Report No. 51, Department of Statistics, Stanford University.

Belloni, A., Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli* 19:521–547.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* 37:373–384.

Cattaneo, M. D., Jansson, M., Newey, W. K. (2012). Alternative asymptotics and the partially linear model with many regressors. Working Paper, MIT.

Donoho, D. L., Johnstone, I. M. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika* 81:425–455.

Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96:1348–1360.

James, W., Stein, C. M. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. 1:361–380.

Lehmann, E. L., Casella, G. (1998). *Theory of Point Estimation*. 2nd ed. New York: Springer.

Pinsker, M. S. (1980). Optimal filtration of square-integrable signals in Gaussian white noise. *Problems of Information Transmission* 16:120–133.

Pötscher, B. M., Leeb, H. (2008). Sparse estimators and the oracle property, or the return of Hodges' estimator. *Journal of Econometrics* 142:201–211.

Pötscher, B. M., Leeb, H. (2009). On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *Journal of Multivariate Analysis* 100:2065–2082.

Pötscher, B. M., Schneider, U. (2009). On the distribution of the adaptive LASSO estimator. *Journal of Statistical Planning and Inference* 139:2775–2790.

Stein, C. M. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Statist. Probab.* 1:197–206.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Annals of Statistics* 9:1135–1151.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B* 58:267–288.

van der Geer, S., Bühlmann, P. (2011). *Statistics for High-Dimensional Data*. New York: Springer

Wasserman, L. (2006). *All of Nonparametric Statistics*. New York: Springer.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 101:1418–1429.

Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B* 67:301–320.