Statistica Sinica **26** (2016), 35-67
doi:http://dx.doi.org/10.5705/ss.2014.042

A STUDY OF ERROR VARIANCE ESTIMATION IN LASSO REGRESSION

Stephen Reid, Robert Tibshirani and Jerome Friedman

Stanford University

Abstract: Variance estimation in the linear model when $p > n$ is a difficult problem. Standard least squares estimation techniques do not apply. Several variance estimators have been proposed in the literature, all with accompanying asymptotic results proving consistency and asymptotic normality under a variety of assumptions.

It is found, however, that most of these estimators suffer large biases in finite samples when true underlying signals become less sparse with larger per element signal strength. One estimator seems to merit more attention than it has received in the literature: a residual sum of squares based estimator using Lasso coefficients with regularisation parameter selected adaptively (via cross-validation).

In this paper, we review several variance estimators and perform a reasonably extensive simulation study in an attempt to compare their finite sample performance. It would seem from the results that variance estimators with adaptively chosen regularisation parameters perform admirably over a broad range of sparsity and signal strength settings. Finally, some intial theoretical analyses pertaining to these types of estimators are proposed and developed.

Key words and phrases: Cross-validation, error variance estimation, lasso.

1. Introduction

Consider the linear model

$$Y = X\beta + \epsilon,$$

where Y is an n-vector of independently distributed responses, X an $n \times p$ matrix with individual specific covariate vectors as its rows and ϵ an n-vector of i.i.d. random variables (usually assumed Gaussian) each with mean 0 and variance σ^2.

When $p > n$, one cannot estimate the unknown coefficient vector β uniquely via standard least squares methodology. In fact, it is probably ill-advised to use least squares to estimate the vector even when $p \leq n$ with p close to n, since standard errors are likely to be high and parameter estimates unstable. In this instance, if one can assume that β is reasonably sparse with many zero entries,

a successful method for selecting the nonzero elements of β and estimating them is the Lasso estimator proposed by Tibshirani (1996), obtained by minimising

$$\frac{1}{2}||Y - X\beta||_2^2 + \lambda||\beta||_1,$$

where the parameter λ is predetermined and controls the amount of regularisation. Broadly speaking, the higher the value of λ, the more elements of the estimated β vector are set to 0 and the more the nonzero entries are shrunken toward 0. Smaller λ implies less regularisation and more nonzero β with larger (absolute) coefficients. The number of non-zero coefficients is not monotone in the value of λ, since sometimes we have to delete variables as we decrease λ, especially at smaller values. However, the notion of smaller λ meaning "less regularisation" is a good rule of thumb.

Much has been written about the model selection and prediction properties of this class of estimators, but it is only recently that people have turned to developing significance tests for the estimated coefficients. Examples include Lockhart et al. (2013) and Javanmard and Montanari (2013). Each of these requires a good estimate of the error variance σ^2 to plug into their chosen test statistics. The problem of estimating error variance when $p > n$ is interesting in its own right and several estimators have been proposed by different authors.

The aim of this paper is to review some of these estimators and to run a comprehensive simulation comparison of their estimation performance over a broad range of parameter vector sparsity and signal strength settings. Perhaps such a comprehensive simulation comparison may reveal the most promising estimator, helping to guide research into fruitful directions. In particular, a promising estimator seems to be

$$\hat{\sigma}^2 = \frac{1}{n - \hat{s}_{\hat{\lambda}}}||Y - X\hat{\beta}_{\hat{\lambda}}||_2^2,$$

where $\hat{\beta}_\lambda$ is the Lasso estimate at regularisation parameter λ, $\hat{\lambda}$ is selected via cross-validation and $\hat{s}_{\hat{\lambda}}$ is the number of nonzero elements in $\hat{\beta}_{\hat{\lambda}}$.

2. Review of Error Variance Estimators

In this section, we review some of the error variance estimators proposed recently and list some of their theoretical properties, as well as the assumptions under which these properties hold.

2.1. The oracle

The ideal variance estimator is the oracle estimator:

$$\hat{\sigma}_O^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'\beta^*)^2, \tag{2.1}$$

where β^* is the true (unknown) coefficient vector with s nonzero elements. This estimator (times n) has a χ^2 distribution with n degrees of freedom and serves as a sample variance for the zero mean ϵ. Obviously this is not a viable estimator in practice, because we do not know β^*. However, it is useful for comparison purposes in a simulation study.

2.2. Residual sum of squares based estimators

Fan, Guo, and Hao (2012) consider estimators of the form

$$\hat{\sigma}^2_{L,\lambda_n} = \frac{1}{n - \hat{s}_{L,\lambda_n}} \sum_{i=1}^{n} (Y_i - X_i' \hat{\beta}_{\lambda_n})^2, \tag{2.2}$$

where β_λ is the Lasso coefficient vector estimate, and $\hat{s}_{L,\lambda}$ the number of nonzero elements of this vector, at regularisation parameter λ. Greenshtein and Ritov (2004) show estimators of this form to be consistent for σ^2 under some technical conditions on the population moments of Y and X. Consistency holds if $\lambda_n = O(\sqrt{\log(p)/n})$.

Fan, Guo, and Hao (2012) show that this estimator has a limiting zero mean normal distribution as $n \to \infty$, $s \log(p)/\sqrt{n} \to 0$ and $\lambda_n \propto \sigma\sqrt{\log(p)/n}$. Furthermore, this limiting distribution has the same variance as the asymptotic variance of the oracle estimator.

Their results are gleaned by making assumptions on the elements of matrix X (assumed to be bounded absolutely) and the so-called *sparse-eigenvalues*. The smallest and largest sparse eigenvalues are defined as

$$\phi_{min}(m) = \min_{M:|M|\leq m} \lambda_{min}(\frac{1}{n} X_M^T X_M),$$

and

$$\phi_{max}(m) = \max_{M:|M|\leq m} \lambda_{max}(\frac{1}{n} X_M^T X_M),$$

where M is a set of integers selected from $\{1, \ldots, p\}$, X_M is the $n \times M$ matrix obtained by selecting columns from X indexed by elements of M, and $\lambda_{min}(A)$ and $\lambda_{max}(A)$ are the smallest and largest eigenvalues of matrix A. Assumptions are made bounding the asymptotic behaviour of these sparse eigenvalues. A lower bound on the smallest sparse eigenvalue seems to be particularly important. These types of assumptions seem to be quite prevalent in the literature that pertains to our problem.

Although heartening, results of this kind are not useful in practice. The choice of λ is very important in the pursuit of an accurate finite sample estimator. Its size controls both the number of variables selected and the degree to which their estimated coeffcients are shrunk to zero. Set λ too large and we do not select

all signal variables, leading to rapidly degrading performance (exhibited mostly by large upward bias) when the true β becomes less sparse with larger signal per element. On the other hand, should we set λ too small, we would select many noise variables, allowing spurious correlation to decrease our variance estimate, leading to substantial downward bias. Simulation results seem to suggest there is a fine balance to be maintained when selecting the appropriate λ.

2.3. Cross-validation based estimators

Considerations around the selection of an appropriate λ lead us inexorably toward an adaptive selection method. In particular, one can define

$$\hat{\sigma}^2_{L,\hat{\lambda}} = \frac{1}{n - \hat{s}_{L,\hat{\lambda}}} \sum_{i=1}^{n} (Y_i - X_i' \hat{\beta}_{\hat{\lambda}})^2, \tag{2.3}$$

where $\hat{\lambda}$ is selected using K-fold cross-validation. K is usually set to 5 or 10. Our simulation results suggest that this estimator is robust to changes in signal sparsity and strength, more so than its competitors. Fan, Guo, and Hao (2012) lament the downward bias of this estimator. They claim that it is affected by spurious correlation. Although this downward bias seems to be borne out in our simulation results, it does not seem too large and stems from a heavy left tail in its empirical distribution. The median estimate tends to be very close to the true σ^2 under a surprisingly broad range of sparsity and signal strength settings.

Very little theory exists detailing the properties of this estimator. Homrighausen and McDonald (2013) prove a result on the persistence of this estimator that can, with a suitable sparsity assumption on the true β, be adapted to a consistency result for an estimate closely resembling $\hat{\sigma}^2_{L,\hat{\lambda}}$.

An implication of their result is that if the true underlying coefficient vector β^* is sufficiently sparse, $||\beta^*||_1 = o((n/\log(n))^{1/4})$, then

$$\frac{n - \hat{s}}{n} \hat{\sigma}^2_{L,\hat{\lambda}} \xrightarrow{P} \sigma^2.$$

If one can assume, as do Fan, Guo, and Hao (2012), that $\hat{s} = o_P(n)$, then $\hat{\sigma}^2_{L,\hat{\lambda}}$ is also consistent. We are not aware of a proof of this for cross-validation though. Nothing is said about the finite sample distribution of this estimator, or whether any asymptotic distribution obtains for that matter.

Fan, Guo, and Hao (2012) propose two other cross-validation based variance estimators. The first defines the K cross-validation folds as $\{D_1, D_2, \ldots, D_K\}$ and computes

$$\hat{\sigma}^2_{CVL} = \min_{\lambda} \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in D_k} (Y_i - X_i' \hat{\beta}^{(-k)}_{\lambda})^2, \tag{2.4}$$

where $\hat{\beta}_{\lambda}^{(-k)}$ is the Lasso estimate at λ over the data after the kth fold is omitted. They try $K = 5, 10$, and n, the latter corresponding to leave-one-out cross-validation. They find in their simulations that the estimate is consistently above the true error variance. We found the same tendency and this estimator is omitted from the simulation study exposition in the next section.

A second estimator uses cross-validation to select the optimal regularisation parameter $\hat{\lambda}$ and then finds the set of indices corresponding to nonzero entries in $\hat{\beta}_{\hat{\lambda}}$. Call this set $\hat{M}$. The "naïve" two-stage Lasso estimator is then defined as

$$\hat{\sigma}_{NL}^2 = \frac{1}{n - |\hat{M}|} ||(I - X_{\hat{M}}(X'_{\hat{M}} X_{\hat{M}})^{-1} X'_{\hat{M}})Y||_2^2. \tag{2.5}$$

This estimator suffers from downward bias for sparse β, because the Lasso tends to overselect (including the vast majority of signal variables and a few noise variables). Least squares estimates of parameters are not shrunk toward zero and inclusion of additional noise variables (that seem well correlated with the response) drives down the variance estimate. Wasserman and Roeder (2009) demonstrate the overselection property of the Lasso. The downward bias of this estimator is made apparent in our simulation results.

2.4. Refitted cross-validation (RCV) estimator

In an attempt to overcome the downward bias caused by spurious correlation in the naïve Lasso estimator, Fan, Guo, and Hao (2012) propose a refitted cross-validation (RCV) estimator. They split the dataset into two (roughly) equal parts $X^{(1)}$ and $X^{(2)}$. On the first part, $X^{(1)}$, they fit the Lasso, using cross-validation to determine the optimal regularisation parameter $\hat{\lambda}_1$ and corresponding set of nonzero indices $\hat{M}_1$. Using those columns in $X^{(2)}$ indexed by $\hat{M}_1$ they obtain the variance estimate

$$\hat{\sigma}_1^2 = \frac{1}{n - |\hat{M}_1|} ||(I - X^{(2)}_{\hat{M}_1}(X^{(2)\prime}_{\hat{M}_1} X^{(2)}_{\hat{M}_1})^{-1} X^{(2)\prime}_{\hat{M}_1})Y||_2^2.$$

They then repeat the mirror image procedure on $X^{(2)}$, obtaining $\hat{\lambda}_2$, $\hat{M}_2$ and $\hat{\sigma}_2^2$. The RCV variance estimate is then obtained as

$$\hat{\sigma}_{RCV}^2 = \frac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}{2}. \tag{2.6}$$

The authors prove consistency and asymptotic normality (with asymptotic variance the same as that of the oracle estimator) of this estimator under slightly weaker conditions than those used for proving similar results for $\hat{\sigma}^2_{L,\lambda_n}$. They argue that breaking up the dataset counters the effect of spurious correlation, since spurious noise variables selected on one half are unlikely to produce significant

least squares parameter estimates on the second half, reducing the negative bias associated with the overselection of the Lasso selector.

Theoretical results aside, the finite sample performance of this estimator seems to suffer when β is less sparse and has larger signal per element. The plug-in Lasso estimator $\hat{\sigma}^2_{L,\hat{\lambda}}$ remains anchored around the true σ^2 for a broader array of sparsity and signal strength settings.

2.5. SCAD estimator

The Lasso is just one method for selecting the variables to have nonzero coefficients in our variance estimator. Any other valid variable selection method could be used to estimate error variance in the spirit of $\hat{\sigma}^2_{L,\hat{\lambda}}$. One such method is the Smoothly Clipped Absolute Deviation Penalty (SCAD) of Fan and Li (2001). Instead of using an ℓ_1 penalty, they minimise

$$\frac{1}{2}||Y - X\beta||_2^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|),$$

where $p'_\lambda(\theta) = \lambda\left(I(\theta \leq \lambda) + [(a\lambda - \theta)_+/(a-1)\lambda]I(\theta > \lambda)\right)$ for some $a > 2$ (usually 3.7) and $\theta > 0$. This penalty is chosen for its good model selection properties. Although no longer a convex criterion, the authors claim to have a stable and reliable algorithm for determining the optimal β with good properties. Indeed, their simulations seem to suggest that SCAD outperforms the Lasso at variable selection in the low noise case when both have their regularisation parameters chosen by cross-validation.

Given a method with good variable selection performance (it selects the signal variables and few or none of the noise variables), we have a hope of mimicking an oracle estimator that is privy to the correct β. Fan, Guo, and Hao (2012) define their SCAD variance estimator as:

$$\hat{\sigma}^2_{SCAD} = \frac{1}{n - \hat{s}_{\hat{\lambda}}}||Y - X\hat{\beta}_{SCAD,\hat{\lambda}}||_2^2, \tag{2.7}$$

where $\hat{\beta}_{SCAD,\hat{\lambda}}$ is the SCAD estimate of β at the regularisation parameter $\hat{\lambda}$ selected by cross-validation. Again, consistency and asymptotic normality can be shown for this estimator with an appropriately chosen, deterministic regularisation parameter sequence λ_n. Our simulations suggest that it performs comparably to $\hat{\sigma}^2_{L,\hat{\lambda}}$.

2.6. Scaled sparse linear regression estimators

Stadler, Buhlmann and van der Geer (2010) introduce the notion of estimating jointly the parameter vector and error variance in the context of mixture

regression models. Sun and Zhang (2010) refine this notion for the non-mixture case and explore the properties of this new estimator in Sun and Zhang (2012).

In particular, the latter pair proposes the joint optimisation in (β, σ) of the jointly convex criterion (called the "scaled Lasso" criterion)

$$\frac{||Y - X\beta||_2^2}{2n\sigma} + \frac{\sigma}{2} + \lambda_0||\beta||_1, \tag{2.8}$$

where λ_0 is some predetermined fixed parameter. An iterative, alternating optimisation algorithm is given where, given a current estimate $\hat{\beta}^{current}$, parameter estimates are updated as:

$$\begin{aligned} \hat{\sigma} &= \frac{||Y - X\hat{\beta}^{current}||_2}{\sqrt{n}}, \\ \lambda &= \hat{\sigma}\lambda_0, \\ \hat{\beta}^{current} &= \hat{\beta}_\lambda, \end{aligned}$$

where $\hat{\beta}_\lambda$ is the Lasso estimate of β at regularisation parameter λ. These steps are interated until the parameter estimates converge.

The authors go on to show consistency, asymptotic normality and oracle inequalities for this estimator under a *compatibility* assumption (detailed in their paper) and assumptions on the sparse eigenvalues of X. The finite sample success of this method, however, hinges on the choice of λ_0. The asymptotic results hold when $\lambda_0 \propto \sqrt{\log(p)/n}$, but finite sample accuracy will depend greatly on an appropriate choice of the proportionality constant. Simulation results from their paper suggest that $\sqrt{2}$ is a good choice for the proportionality constant, but our simulation results show rapid degradation as the true β becomes less sparse with larger per element signal.

Another estimator proposed by Sun and Zhang (2012) uses the scaled Lasso criterion to find $\hat{M}_{SZ}$ - the set of indices corresponding to nonzero $\hat{\beta}^{current}$ after the final iteration. Once obtained, another estimator is defined as

$$\hat{\sigma}^2_{SZLS} = \frac{1}{n - |\hat{M}_{SZ}|}||(I - X_{\hat{M}_{SZ}}(X'_{\hat{M}_{SZ}}X_{\hat{M}_{SZ}})^{-1}X'_{\hat{M}_{SZ}})Y||_2^2. \tag{2.9}$$

The authors tout the finite sample accuracy of this estimator.

In a recent paper, Sun and Zhang (2013) propose a different value for λ_0. With this value, tighter error bounds are achieved than in their previous paper. In particular, they propose

$$\lambda_0 = \sqrt{2}L_n(\frac{k}{p}), \tag{2.10}$$

with $L_n(t) = \Phi^{-1}(1-t)/\sqrt{n}$, where Φ is the standard Gaussian cdf and k is the solution to

$$k = L_1^4(\frac{k}{p}) + 2L_1^2(\frac{k}{p}).$$

A least squares after scaled Lasso estimator (as in (2.9)) is also proposed for this level of the regularisation parameter. All four of the scaled Lasso estimators were included in our simulation study.

2.7. Method of moments estimators

Dicker (2014) takes a different tack. Instead of attempting to emulate the sum of squares estimator of standard least squares regression methodology, he makes distributional assumptions on both the errors ϵ and the columns of the predictor matrix X.

He retains the standard assumption that $\epsilon \sim N_n(0, \sigma^2 I_n)$, although he makes it at the outset; the subsquent derivation of his estimator depending heavily on this assumption. Furthermore, he assumes that each of the n rows of X (call the ith one x_i) is normally distributed: $x_i \sim N_p(0, \Sigma)$. Also, all ϵ_i and x_i are assumed independent.

These distributional assumptions allow one to compute the expectations of the quantities $||y||^2$ and $||X'y||^2$. Equating these moments to their sample counterparts enables one to derive estimators for σ^2.

He proposes two estimators. The first holds when we assume $\Sigma = I_p$,

$$\hat{\sigma}^2_{D1} = \frac{p+n+1}{n(n+1)}||y||^2 - \frac{1}{n(n+1)}||X'y||^2, \tag{2.11}$$

while a second estimator is an approximate method of moments estimator for the case of general Σ,

$$\hat{\sigma}^2_{D2} = \left[1 + \frac{p\hat{m}_1^2}{(n+1)\hat{m}_2}\right]\frac{1}{n}||y||^2 - \frac{\hat{m}_1}{n(n+1)\hat{m}_2}||X'y||^2, \tag{2.12}$$

where

$$\hat{m}_1 = \frac{1}{p}\text{tr}\left(\frac{1}{n}X'X\right), \hat{m}_2 = \frac{1}{p}\text{tr}\left[\left(\frac{1}{n}X'X\right)^2\right] - \frac{1}{pn}\left[\text{tr}\left(\frac{1}{n}X'X\right)\right]^2.$$

He shows how these estimators are consistent for σ^2 and have asymptotic Gaussian distributions.

3. A Simulation Study

The merits of each of the estimators mentioned are demonstrated by the authors who devised them. Asymptotic results are gleaned for each and simulation studies run to show some applicability. In this section we exact upon the entire collection a fairly extensive simulation study. In the study we control the sparsity of the underlying true β vector as well as its signal-to-noise ratio (SNR). The correlation between columns of the X matrix is also controlled. The aim of

the study is to reveal the strengths and weaknesses of the estimators (and the sparsity-signal strength combinations in which these are most clearly revealed). In particular, we would like to ascertain which estimator provides reasonable estimates of the error variance over the broadest range of sparsity and signal strength settings.

Use of the Lasso (and other sparsity-inducing coefficient estimators) makes a large bet on sparsity. Most of the good results obtained for this class of estimators make some crucial assumptions about the sparsity of the underlying ground truth. The variance estimators at issue also rely heavily on the notion of finding *the* small set of nonzero coefficients and using it to remove the signal from the response, leaving only random error, the variance of which can then be obtained. In practice though, we are rarely certain about the extent of sparsity of the ground truth. A variance estimator that performs reasonably over a broad range of ground truth settings lends some peace of mind.

3.1. Simulation parameters

All simulations were run at a sample size of $n = 100$. For the number of total predictors we took $p = 100,\ 200,\ 500,\ 1{,}000$. Elements of the predictor matrix X were generated randomly as $X_{ij} \sim N(0,1)$. Two correlation structures for columns of X were considered, each parameterised by parameter ρ: $\mathrm{Cor}(X_i, X_j) = \rho\ \forall i, j$; $\mathrm{Cor}(X_i, X_j) = \rho^{|i-j|}\ \forall i, j$.

The true β was generated in steps. First, the number of nonzero elements was set to $p_{nz} = \lceil n^{\alpha} \rceil$. The parameter α controls the degree of sparsity of β: the higher the α; the *less* sparse the β. It ranges between 0 and 1, except when we set it to $-\infty$ to enforce $\beta = 0$. The indices corresponding to nonzero β were then selected randomly. Their values were set equal to that of a random sample from a $Laplace(1)$ distibution. The elements of the resulting β were then scaled such that the signal-to-noise ratio, defined as $\beta'\Sigma\beta/\sigma^2$, was some predetermined value, snr. Here Σ is the covariance matrix of the elements of a single row of X.

Simulations were run over a grid of values for each parameter. In particular, $\rho = 0, 0.2, 0.4, 0.6, 0.8$; $\alpha = 0.1, 0.3, 0.5, 0.7, 0.9$ and $snr = 0.5, 1, 2, 5, 10, 20$. At each setting of the parameters, $B = 100$ replications of each of a collection of error variance estimators were obtained. The collection of estimators considered were as follows.

- The oracle estimator at (2.1).
- The cross-validation based Lasso estimator $\hat{\sigma}^2_{L,\hat{\lambda}}$ at (2.3), denoted *CV_L* in the simulation output.
- The naïve Lasso estimator $\hat{\sigma}^2_{NL}$ at (2.5), denoted *CV_LS*.
- The SCAD estimator $\hat{\sigma}^2_{SCAD}$ at (2.7), denoted *CV_SCAD*.

- The RCV estimator $\hat{\sigma}^2_{RCV}$ at (2.6).
- The scaled Lasso estimator of Sun and Zhang (2012) at (2.8), denoted *SZ* in output.
- Its least-squares-after-scaled-Lasso version at (2.9), denoted *SZ_LS*.
- The scaled Lasso estimator of Sun and Zhang (2013) with smaller regularisation parameter at (2.10), denoted *SZ2*.
- Its least-squares-after-scaled-Lasso version, denoted *SZ2_LS*.
- The method of moments estimators of Dicker (2014) at (2.11) and (2.12), denoted *D1* and *D2*, respectively.

All figures, tables and results are quoted in terms of the standard deviation estimate. Results were obtained for true error variance values $\sigma = 1, 3$.

3.2. No signal case: $\beta = 0$

We first considered the edge case $\alpha = -\infty$, forcing $\beta = 0$. The *snr* is irrelevant here, because we have no signal. Figure 1 shows boxplots of the replications of the standard deviation estimates when $\rho = 0$ and $\sigma = 1$. True σ is indicated by the horizontal line, for reference.

It is apparent that the CV_L, CV_SCAD, and SZ2 estimators are slightly downward biased, whereas the RCV, SZ, and SZ_LS estimators appear unbiased. The least-squares-after-Lasso-CV estimator (CV_LS) has considerable downward bias (as in SZ2_LS). This probably stems from the tendency of Lasso to overselect when the regularisation parameter is chosen via CV. The relatively large set of predictors chosen, coupled with the ill effects of spurious correlation when estimating via least squares, probably contribute most significantly to this downward bias.

The method of moments estimators tend to be median unbiased, but their variances increase considerably as p increases relative to n. This increase in variance is most pronounced in the bottom right panel (p = 1,000), where the method of moments estimators have significantly larger variance than the rest.

Median biases for each of the estimators are tabulated in Table 1 for the different n-p combinations. This is defined as $\text{median}_{b=1,2,\ldots,B}\{\hat{\sigma}_b\} - \sigma$, where $\hat{\sigma}_b$ is the bth replication of the standard deviation estimate of interest. It would seem that median biases for CV_L, CV_LS, and CV_SCAD increase (absolutely) as p increases. Although lamentable, the biases of CV_L and CV_SCAD are not that large, particularly when compared to the biases of the other estimators when we start increasing the signal (see below). Furthermore, in practice, the assumption is often that there is indeed a signal. This is usually the point of a study. This particular setup then, may not be encountered too often in practice.

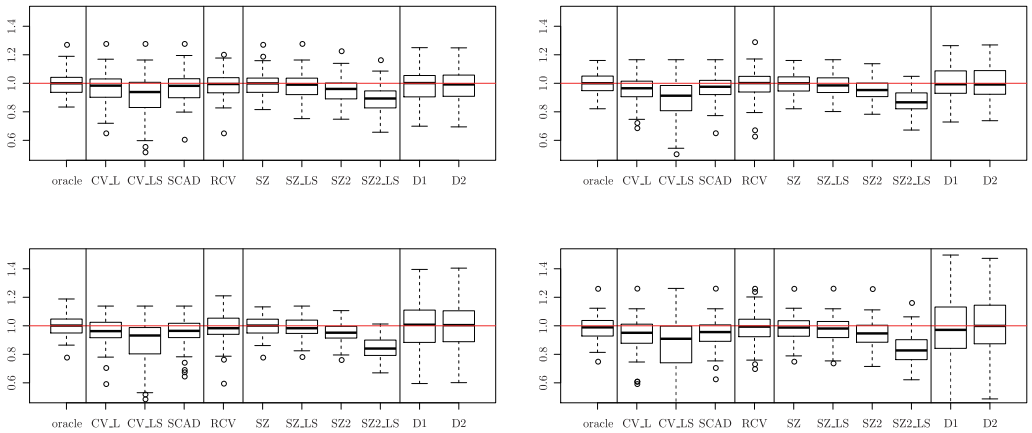

Figure 1. Standard deviation estimates for $\beta = 0$ case. Sample size $n = 100$, predictors $p = 100$, 200, 500, 1,000, moving left to right along rows. $\rho = 0$.

Table 1. Median biases of standard deviation estimators. No signal, $\sigma = 1$, $\rho = 0$.

Method	$p = 100$	$p = 200$	$p = 500$	$p = 1{,}000$
Oracle	0.0000	0.0004	0.0015	-0.0110
CV_L	-0.0165	-0.0348	-0.0374	-0.0491
CV_LS	-0.0612	-0.0871	-0.0680	-0.0910
CV_SCAD	-0.0177	-0.0242	-0.0355	-0.0440
RCV	-0.0050	0.0014	-0.0170	-0.0059
SZ	-0.0006	0.0002	0.0015	-0.0122
SZ_LS	-0.0096	-0.0150	-0.0171	-0.0191
SZ2	-0.0396	-0.0474	-0.0484	-0.0534
SZ2_LS	-0.1065	-0.1332	-0.1596	-0.1727
D1	0.0150	-0.0068	0.0074	-0.0286
D2	-0.0079	-0.0084	0.0048	-0.0008

We also notice from Figure 1 the tight clustering of the estimates around the true σ. None of the clusterings are as tight as that of the oracle, but on the whole, all the standard deviation estimates seem to have low variance (except for CV_LS, D1, and D2). CV_L and RCV seem to produce rare outlier estimates, with those from CV_L always coming in below the true σ. The distribution of CV_L appears skewed to the left, which may make it difficult to analyse, particularly when one wants to ascertain the distribution of a test statistic using this variance estimator. The effort may be merited though, as we will see below that this estimator performs admirably over a broad range of sparsity and signal strength assumptions.

Correlation between the columns of the predictor matrix seems to have little effect on the performance of each of our estimators. Curves (not shown) depicting

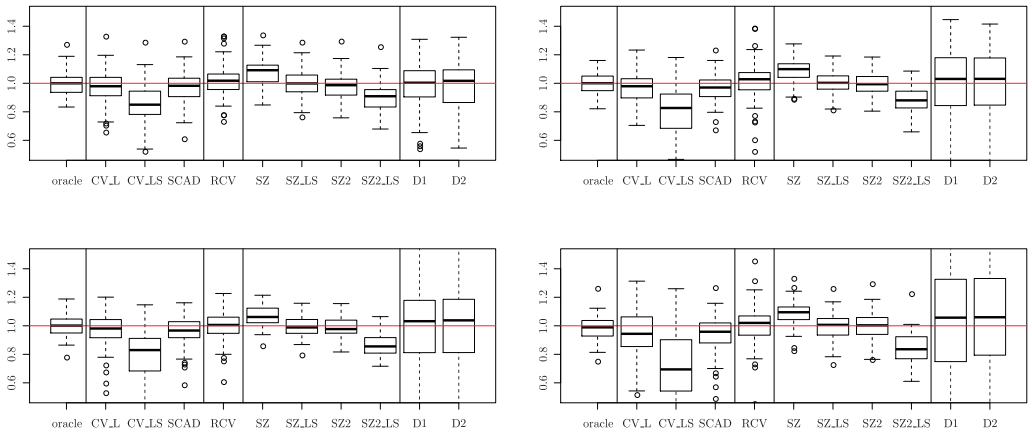

Figure 2. Standard deviation estimates for $\alpha = 0.1$ (sparse). Sample size $n = 100$, predictors $p = 100$, 200, 500, 1,000, moving left to right along rows. $\rho = 0$, $snr = 1$.

median standard deviation estimates as a function of predictor correlation ρ all seem relatively flat, with the estimators retaining their properties as discussed above. Similar looking curves are obtained for the high noise case, $\sigma = 3$ (also not shown).

3.3. Effect of sparsity: changing α

The true value of the CV_L estimator becomes apparent once we consider different sparsity levels and signal strength settings. It should be noted that each of the estimators eventually breaks down when signals become non-sparse and large. This is reflected by the very large upward biases in all of the estimators. The question then is not whether we can find a silver bullet for all conceivable ground truths, but rather one that performs reasonably for a broad range of possible ground truths.

Our first consideration in the quest for such a broadly applicable estimator is the effect of decreased sparsity. In our simulation, the sparsity level is controlled by changing the value of α. The higher the α; the less sparse the ground truth β becomes.

Figure 2 shows the boxplots of standard deviation estimates when $\alpha = 0.1$ and $snr = 1$ in the uncorrelated case ($\rho = 0$). Notice that the median bias of the CV_L estimator seems to have decreased, while that of the CV_SCAD estimator has remained negative, roughly of the same size as in the no-signal case. Downward bias in CV_LS now seems more pronounced, while the SZ estimator has become upwardly biased, with bias increasing with p. SZ_LS performs best,

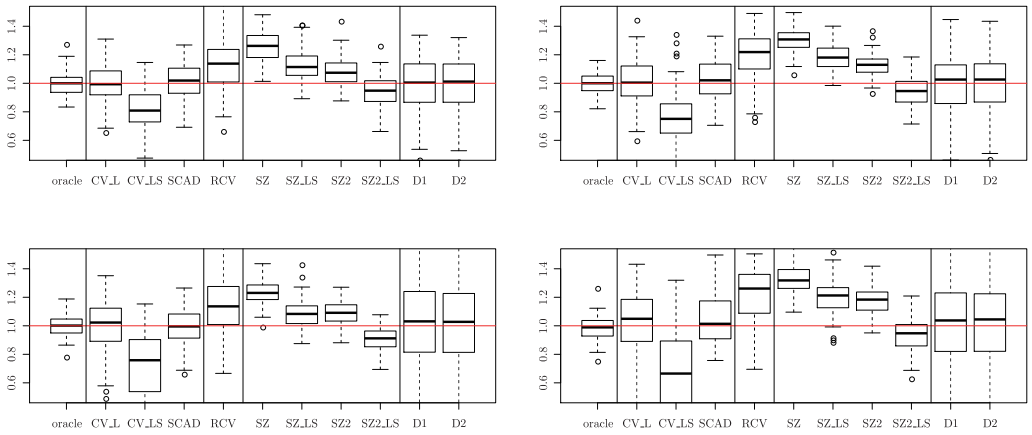

Figure 3. Standard deviation estimates for $\alpha = 0.5$ (less sparse). Sample size $n = 100$, predictors $p = 100$, 200, 500, 1,000, moving left to right along rows. $\rho = 0$, $snr = 1$.

being unbiased with a tight distribution around its median. CV_L and RCV perform comparably.

This changes quite dramatically when we set $\alpha = 0.5$, as in Figure 3. Here we see that the CV_L and CV_SCAD estimators are the only two estimators without substantial biases in either direction. RCV, SZ, and SZ_LS have all become biased upward by roughly 20%, while SZ2 becomes increasingly more upwardly biased as p increases, starting with a bias of about 7.5% at $p = 100$, growing to about 18% when $p = 1{,}000$.

Figures 4 and 5 were generated under positive correlation $\rho = 0.4$. The former shows output for the case where all pairwise correlations are $\rho = 0.4$, while the latter has the decaying correlation structure described earlier.

In practice, one would expect the p predictors to be correlated. These two plots are meant to reveal estimator performance under more realistic assumptions than the $\rho = 0$ assumption of Figure 3. Notice that results are qualitatively similar over Figures 3, 4, and 5, albeit slightly more muted for larger p when $\rho = 0.4$.

3.3.1. Explaining the biases

In attempt to understand why the biases obtain, consider the oracle estimator

$$\hat{\sigma}_O^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'\beta^*)^2.$$

The success of this estimator hinges on the fact that it knows the true β (which we call β^*). It is able to remove all the signal from the observed Y_i, leaving only the errors ϵ_i, the variance of which we wish to measure.

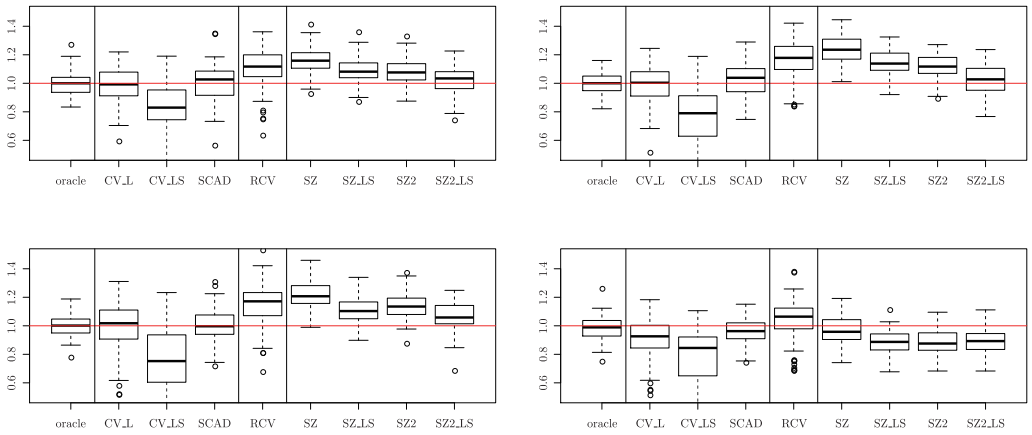

Figure 4. Standard deviation estimates for $\alpha = 0.5$ (less sparse). Sample size $n = 100$, predictors $p = 100$, 200, 500, 1,000, moving left to right along rows. $\rho = 0.4$ (all pairwise), $snr = 1$.

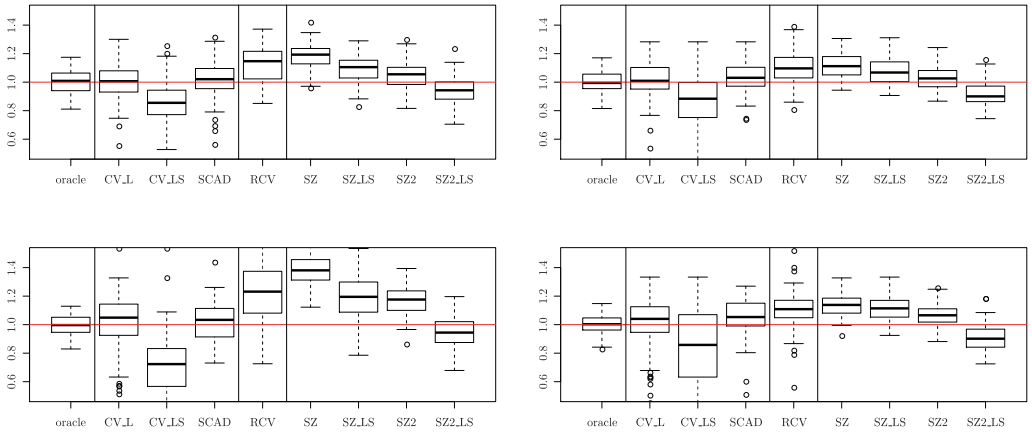

Figure 5. Standard deviation estimates for $\alpha = 0.5$ (less sparse). Sample size $n = 100$, predictors $p = 100$, 200, 500, 1,000, moving left to right along rows. $\rho = 0.4$ (decaying), $snr = 1$.

Other estimators (except the method of moment estimators) attempt to emulate the form of the oracle, but none of them are privy even to the set of non-zero β_j, let alone their true values. Each of these estimators needs to estimate the set of non-zero estimators and then place values on their coefficients. Departures from oracle performance occur when true signal variables are not selected (false negatives), irrelevant variables are selected (false positives) and when estimates of the coefficient values do not match their true underlying values.

Figure 6 is a diagnostic plot showing three measures pertaining to the quality of the estimated β for each of the methods (CV Lasso, CV_SCAD, SZ, SZ2, and

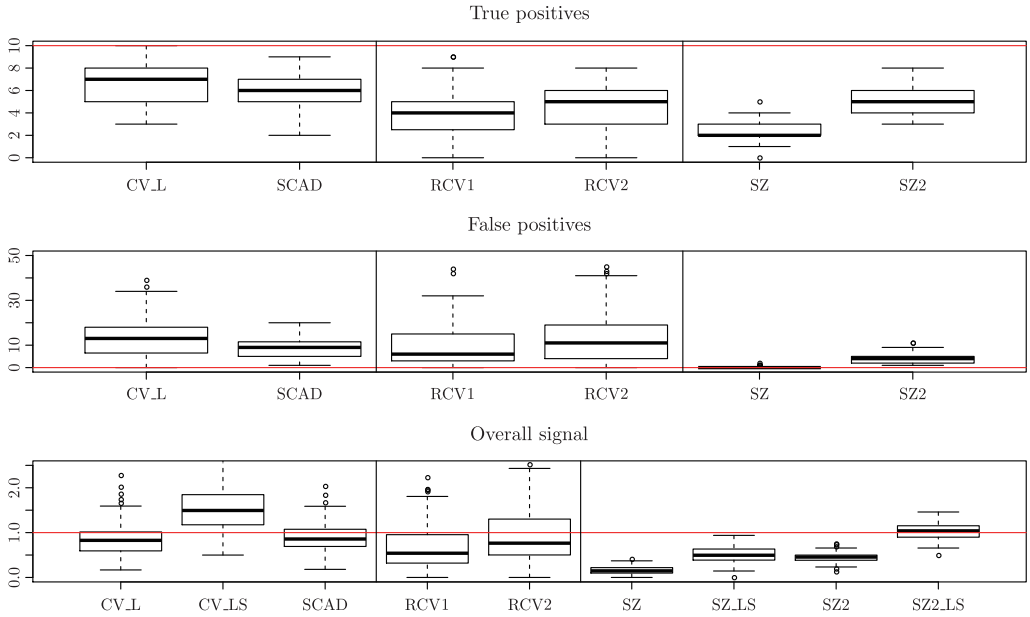

Figure 6. Diagnostic plot. Top panel shows boxplots of the number of non-zero coefficients correctly identified by each procedure over the $B = 100$ simulation runs. Horizontal line at 10 – the true number of signal variables. Middle panel shows the number of zero coefficients incorrectly given non-zero values. Horizontal line at zero – the target number of false positives. Bottom panel shows the ratio of the estimated signal to the true signal $\sum_{j=1}^{p} |\hat{\beta}_j| / \sum_{j=1}^{p} |\beta_j^*|$. Horizontal line at 1, reflecting the ratio that would occur should we capture the true signal perfectly. $p = 100$, $\alpha = 0.5$ (so that there are 10 signal variables and 90 zero variables), and $snr = 1$.

both halves of the RCV - labelled RCV1 and RCV2). Parameters for this figure are $p = 100$, $\alpha = 0.5$ and $snr = 1$. This is one of many such figures that can be drawn, but this one is representative.

CV_L and CV_SCAD tend to select more of the signal variables than do the other variables. None of the methods select all the signal variables. This leads to considerable upward bias as signal size increases, as the residual sum of squares on which all of these estimators are based would inflate with the signal not successfully removed from it.

CV_L and CV_SCAD seem to counter this shortcoming by selecting a moderate number of irrelevant variables and giving them non-zero coefficients (middle). The balance between missing signal variables and capturing irrelevant variables seems to lead to an estimated coefficient vector with signal size rather close to that of the true parameter vector (bottom panel).

RCV seems to select too few signal variables, making it difficult to strike a

balance to find a decent variance estimate. SZ and SZ2 select fewer true signal variables than CV_L and CV_SCAD and detect almost no false positives. The signal variables not selected by the SZ and SZ2 estimators then degrade their performance as signal size increases. Neither RCV, SZ, nor SZ2 produces signal sizes large enough to match the underlying signal (bottom panel). Least squares estimates tend to have signals larger than the true signal, because they have the same set of nonzero coefficients as their penalised counterparts, but with larger, unpenalised coefficients.

Large upward biases occur because none of the methods select all the true signal variables. Some methods seem to find a balance between selecting signal and non-signal variables to produce reasonably good variance estimates over a broad range of sparsity and signal size settings. Why CV_L and CV_SCAD behave this way is not fully understood, but we suspect that the adaptive selection of the regularisation parameter contributes.

In particular, it would seem that adaptively chosen regularisation parameters are less onerous in their omitting of variables and the biasing of non-zero coefficients. Although the numerical comparison of the regularisation parameters chosen by the CV, SCAD, SZ, and SZ2 methods is probably not sensible, one gets the impression that those chosen by CV and SCAD are in some sense "smaller" than those selected by the SZ methods, leading to more, less biased, non-zero coefficients in their fits. The SZ methods tend to choose their regularisation parameters large enough so as to guarantee control of the noise variables (to ensure asymptotically that they are not selected). This focus on noise could lead to reduced control over signal in small samples.

The residual sum of square type variance estimators rely, at least in principle, on a good estimate of the underlying signal. The estimated signal is removed from the response leaving, hopefully, only the noise, the average of which provides a good estimate of noise variance. A direct measure of the quality of signal estimation is

$$\frac{1}{n}||X\hat{\beta} - X\beta^*||_2^2,$$

where $\hat{\beta}$ is the estimated coefficient vector for the method at hand and β^* is the true underlying signal.

Table 2 shows medians of this average signal bias for the CV_L and SZ methods (and their least squares equivalents). That of CV_SCAD is also shown. The median is taken over the $B = 100$ replications at the setting $\alpha = 0.5$, $snr = 1$, and $\rho = 0$. These methods were chosen because they seem to produce the best variance estimates.

CV_L and CV_SCAD tend to produce better signal estimates than their SZ counterparts (lower median signal biases). This could help to explain why the former methods produce better variance estimates.

Table 2. Median over $B = 100$ replications of average signal biases for selected estimation methods. $\alpha = 0.5$, $snr = 1$, $\sigma = 1$, $\rho = 0$.

Method	$p = 100$	$p = 200$	$p = 500$	$p = 1{,}000$
CV_L	0.3269	0.3037	0.3688	0.4465
CV_LS	0.5246	0.5527	0.6806	0.7225
CV_SCAD	0.3261	0.2481	0.3073	0.3998
SZ	0.7326	0.4764	0.6443	0.8079
SZ_LS	0.4951	0.2967	0.3869	0.5864
SZ2	0.6739	0.3115	0.4009	0.4858
SZ2_LS	0.3701	0.3869	0.4460	0.5241

SZ least squares methods sometimes improve on their unadjusted counterparts, while the least squares variant of the CV method produces generally poorer signal estimates. It would seem that the elimination of the coefficient biasing effect of the unadjusted SZ methods helps to counter the bias it experiences in overall signal estimation. Such an effect does not seem to obtain for the CV method.

3.3.2. Different regularization parameters for method CV_L

It is clear from the difference in performance between SZ and SZ2 that the small sample success of a given method hinges crucially on the regularisation parameter chosen with which to apply the method. We have seen that CV_L performs admirably when compared to other methods, but the question begs whether other choices of regularisation parameter – still used in the CV_L estimator – could lead to better variance estimates.

The current CV_L estimate uses the standard mean-squared error CV criterion. We divide the dataset into $K = 10$ folds, fit the lasso at a sequence of λ to $K = 10$ versions of the dataset, each time omitting a different fold, find the mean squared prediction error for each λ over each left-out fold, average and then find that λ minimising the average mean squared error over folds.

Yu and Feng (2013) suggest a modified criterion to be used in each fold. They have two versions: the Exactly Modified Cross-validation Criterion (EMCC) and the Modified Cross-validation Criterion. The claim is that traditional CV method overselects variables, obviously curtailing model selection effectiveness, and reduces prediction accuracy. Their adjustments are meant to lead to regularisation parameters that deliver better model selection and prediction performance. In general, the regularisation parameter chosen here is larger than for the traditional CV criterion, reducing the number of false positives amongst the selected variables. The reader is referred to their paper for details.

For variance estimation, false positives seem, counterintuitively, to aid the endeavor, especially in a setting where we cannot hope to identify exactly the

set of true non-zero coefficients. Modified cross-validation criteria might reduce false positives, which is better for model selection (and perhaps prediction), but this may lead them to have poorer variance estimates. We compared variance estimation performance of the traditional CV method (CV_L) to the same estimator using a regularisation parameter chosen by MCC. We elected to consider this method for its ease of implementation and reasonable variance estimation performance, denoting it by CV_MCC.

A further tweak to the setting of the value of the regularisation parameter comes from a subtlety associated with the definition of the objective. Consider two formulations of the lasso criterion:

$$\frac{1}{2}||y - X\beta||_2^2 + \lambda_n||\beta||_1, \tag{3.1}$$

$$\frac{1}{2n}||y - X\beta||_2^2 + \tau_n||\beta||_1. \tag{3.2}$$

When we use K-fold CV, the sample size is $m = (1-1/K)n$ and the estimated regularisation parameters satisfy $\hat{\lambda}_m = m\hat{\tau}_m$. Since we fix a sequence of λ (or τ) at which we fit the lasso before cross-validation (based on the entire dataset), we have, after cross-validation that

$$n\hat{\tau}_n = n\hat{\tau}_m = \frac{n}{m}\hat{\lambda}_m = \frac{n}{m}\hat{\lambda}_n > \hat{\lambda}_n,$$

so that the criterion in (3.2) leads to a slightly larger sequence of regularisation parameters (and, one supposes, a larger optimal regularisation parameter). For our method, $K = 10$ implies that $m = 0.9n$, which is unlikely to cause much difference. Still, we performed cross validation under (3.1) and (3.2) and compared their results. The former is our original CV_L estimate, while the latter is denoted CV_L_ADJ.

Figure 7 plots the standard deviation estimates over $B = 100$ replicates at the setting $\alpha = 0.5$, $\rho = 0$, and $snr = 1$. There is little or no difference between the performance of CV_L and CV_L_ADJ. CV_MCC seems to produce slightly more upwardly biased estimates, with the bias increasing as we increase p. Again, it seems that CV_L strikes a good balance between true and false positives and the coefficient values assigned to them. The MCC method, touted for its ability to reduce the number of false positives selected, seems to go too far, biasing the remaining selected variables too much for a decent variance estimate.

We thank an anonymous reviewer for the input that led to the development of this section.

3.3.3. Ranging over different α

Figure 8 plots median standard deviation estimates over different values of α. Here we set $snr = 1$ and $\sigma = 1$. Notice how CV_L and CV_SCAD resist

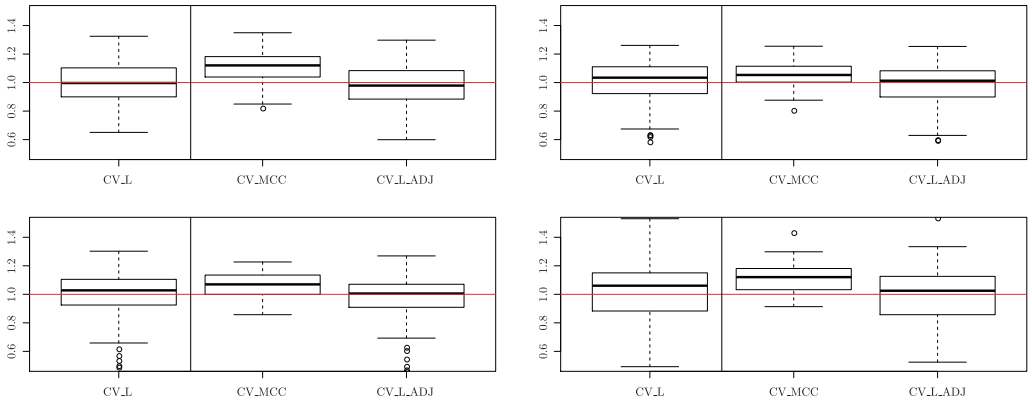

Figure 7. Standard deviation estimates for $\alpha = 0.5$ (less sparse) at different regularisation parameter levels for CV_L estimator. Sample size $n = 100$, predictors $p = 100$, 200, 500, 1,000, moving left to right along rows. $\rho = 0$, $snr = 1$.

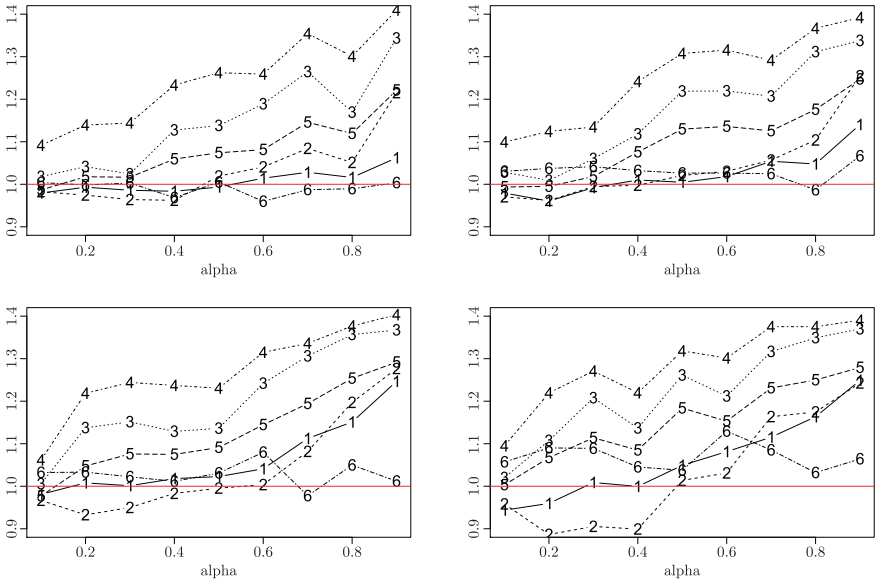

Figure 8. Median standard deviation estimates over different levels of β sparsity. Plot numbers refer to CV_L (1), CV_SCAD (2), RCV (3), SZ (4), SZ2 (5) and D1 (6), respectively. $\sigma = 1$.

upward bias over a broader range of sparsity settings, with CV_L performing most admirably for smaller values of p (top row). This is revealed in the figure by lines 1 and 2 hugging the reference line (true σ) quite closely for α up to 0.5, while by this time, all the other curves have diverged significantly.

The median method of moments estimators are largely immune to a decrease

in sparsity (increase in α); this comes at the expense of larger estimator variance, as reflected in the preceding figures.

3.4. Effect of signal-to-noise ratio

There are two components contributing to the size of β: the degree of sparsity and the per element signal size. For a given sparsity level (number of nonzero elements of β), the higher the SNR (as defined earlier), the higher the per element signal strength. We found in our simulations that individual signal sizes have significant impact on the quality of variance estimates.

Figure 9 is a telling demonstration of the superiority of the CV_L and CV_SCAD estimators, those with data dependent, adaptively selected regularisation parameters. Sparsity level is set at $\alpha = 0.5$, a level both theoretically and anecdotally significant. Theoretical results suggest that, at this level of sparsity, all estimators considered are consistent. This asymptotic result is falsely comforting in finite samples. Clearly some of the estimators are significantly upwardly biased when the signal strength increases.

Anecdotally, it seems as though this level of sparsity coincides with a point of deterioriation of our estimators. As the β vector becomes less sparse beyond this point, the performance of all estimators deteriorates rapidly, suggesting that this level is a significant watershed beyond which we have little hope of decent error variance estimates. As we skirt this precarious edge by increasing the per element signal, we see that CV_L and CV_SCAD remain unaffected, while all other candidates suffer significantly. Although not shown, these plots look similar for the high noise ($\sigma = 3$) case.

Interestingly, the least-squares-after-scaled-Lasso estimator with the smaller regularisation parameter (SZ2_LS) seems to perform admirably here as well (not shown). This, however, is an artifact of setting the sparsity level at $\alpha = 0.5$. For all other sparsity levels, this estimator exhibits significant biases in either direction.

3.5. Effect of predictor correlation: changing ρ

It is interesting to note that correlation between predictors seems to come to the rescue of some of the variance estimators considered. Figure 10 again plots median standard deviations, this time as a function of predictor correlation (ρ). Notice how the large upward bias of the RCV, SZ and SZ2 estimators decreases as ρ increases. Unfortunately, the method of moments estimators perform rather poorly as predictor correlation increases, even D2, designed for general predictor correlation structures.

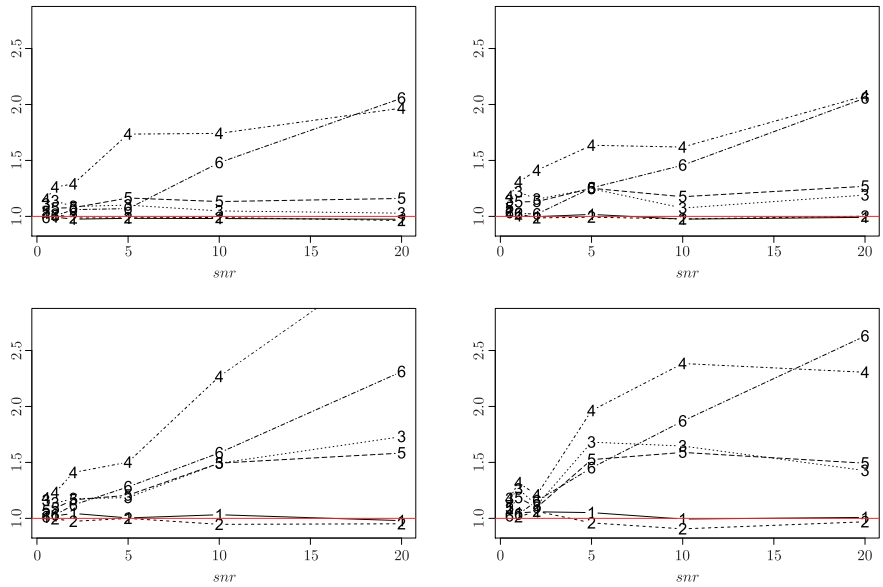

Figure 9. Median standard deviation estimates over different levels of signal-to-noise ratio. Plot numbers refer to CV_L (1), CV_SCAD (2), RCV (3), SZ (4), SZ2 (5), and D1 (6), respectively. $\alpha = 0.5$, $\sigma = 1$.

4. The Effect of σ on the Covariance Test

The previous section detailed the superiority of estimators like CV_L and CV_SCAD, with their adaptively chosen regularisation parameters, in the estimation of error variance when $p \geq n$. Of course, we are rarely interested solely in the estimation of σ. Often its estimation is secondary, with the main goal being to plug it into a test statistic for some hypothesis test. Clearly a badly biased estimate of σ can lead to poor test performance.

In this section, we consider the performance of the covariance test statistic of Lockhart et al. (2013) in a very simple setup. By varying the value used for σ in the denominator of that statistic, we can get an impression of how badly the testing procedure is affected should we have a poor variance estimate.

Consider then, for $n = 100$, an independent sample $Y_i \sim N(\beta_i, \sigma^2)$, for $i = 1, \ldots, n$, $\beta_i = 0$ for $i = 2, \ldots, n$, and some value for β_1 (to be set later). All but one of the sample elements have zero signal. Suppose that the true variance $\sigma^2 = 1$. We want to use the covariance test statistic to test the global null hypothesis $H_0 : \beta_1 = 0$.

Take $|Y|_{(1)} \geq |Y|_{(2)} \geq \dot{\geq} |Y|_{(n)}$ as the order statistics of the absolute sample elements. Lockhart et al. (2013) propose the test statistic:

$$T = \frac{|Y|_{(1)}(|Y|_{(1)} - |Y|_{(2)})}{\sigma^2} = |Y|_{(1)}(|Y|_{(1)} - |Y|_{(2)})$$

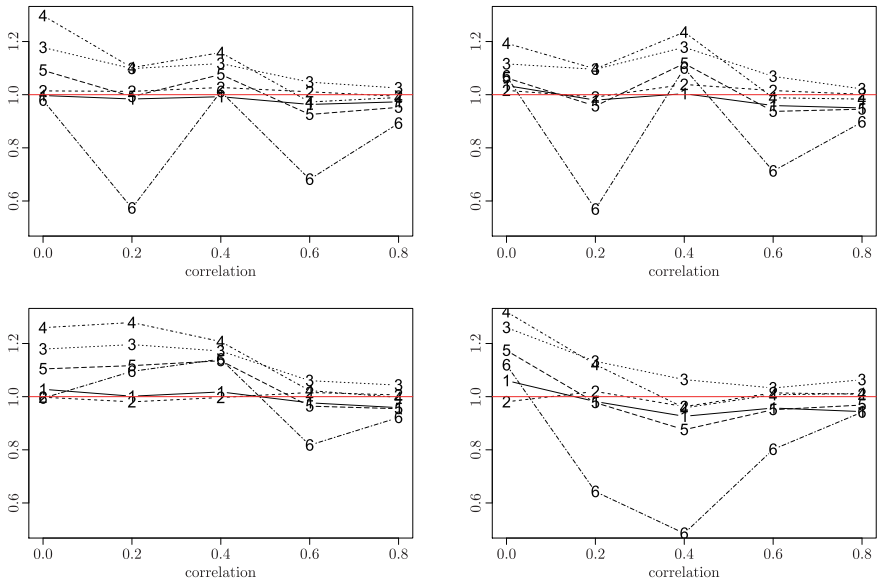

Figure 10. Median standard deviation estimates over different levels of predictor correlation. Sample size $n = 100$ and predictor numbers $p = 100$, 200, 500, 1,000, left to right over rows. Plot numbers refer to CV_L (1), CV_SCAD (2), RCV (3), SZ (4), SZ2 (5) and D2 (6), respectively. $\alpha = 0.5$, $\sigma = 1$, $snr = 1$.

to test H_0. They show that $T \xrightarrow{d} \text{Exp}(1)$ as $n \to \infty$. If we let E_α be such that $P(\text{Exp}(1) > E_\alpha) = \alpha$, then their test suggests that we reject H_0 if $T > E_\alpha$. This gives the test an approximate level of α.

All of their results hinge on the notion that $\sigma = 1$ is known. We performed a small simulation studying the effect of using $\frac{T}{\hat{\sigma}^2}$ instead of T as test statistic for different values of $\hat{\sigma}$. In this way we can get an impression of the asymmetric effects of having $\hat{\sigma} = 0.5$ (downward bias) versus $\hat{\sigma} = 1.5$ (upward bias) and also the extent of the impact on power caused by upward biases.

In our simulation, we considered a sequence of $\hat{\sigma}$: $0.1, 0.2, 0.3, \ldots, 2$. We considered the effect on test size and power, as a function of $\hat{\sigma}$, of using $T/\hat{\sigma}^2$ instead of T in the covariance test. Power was computed under alternatives corresponding to a signal-to-noise ratio of $snr = 0.5, 1, 2, 5, 10, 20$ (this implies $\beta_1 = \sqrt{snr}$).

The left panel Figure 11 shows the estimated size of the test, $P(T/\hat{\sigma}^2 > E_\alpha | \beta_1 = 0)$, as a function of $\hat{\sigma}$. The estimate was obtained be finding the proportion of rejections (at each value of $\hat{\sigma}$) of the test over $B = 1{,}000$ replications

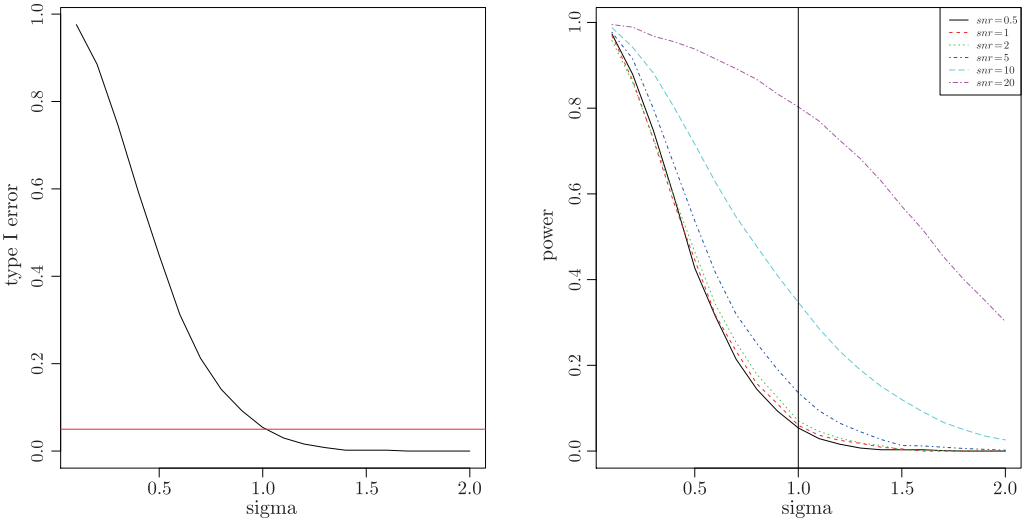

Figure 11. **Left panel:** Estimated size of test at different values of $\hat{\sigma}$. Stated test size is 0.05 (horizontal line for reference). **Right panel:** Estimated power of test at different settings of β_1 as function of $\hat{\sigma}$. Vertical line drawn at $\hat{\sigma} = 1$ for reference.

of the experiment when $\beta_1 = 0$:

$$\hat{size}(\hat{\sigma}) = \frac{1}{B}\sum_{b=1}^{B} I\{\frac{T}{\hat{\sigma}^2} > E_\alpha | \beta_1 = 0\}.$$

The major problem here is encountered when the estimate of σ is downwardly biased, leading to actual test sizes far exceeding the stated size of $\alpha = 0.05$; even moderately downwardly biased variances estimates result in rather large increases in actual test size. For example, for $\hat{\sigma} = 0.9$ and $\hat{\sigma} = 0.8$, we see the actual sizes are 0.092 and 0.141 respectively – quite far removed from the advertised 0.05.

The right hand panel of the same figure shows that upward biases are also quite harmful, in this case, to test power. Each curve represents a different setting of β_1. We compute the average power over $B = 1{,}000$ replications at a given setting of β_1,

$$\hat{power}(\beta_1, \hat{\sigma}) = \frac{1}{B}\sum_{b=1}^{B} I\{\frac{T}{\hat{\sigma}^2} > E_\alpha | \beta_1\}.$$

Larger signals (large β_1) lead to larger power, as expected, but notice the steep decrease in power once we move beyond $\hat{\sigma} = 1$ into the realm of upwardly biased variance estimates. The effect is perhaps better illustrated in Figure 12, which plots the ratio

$$\frac{\hat{power}(\beta_1, \hat{\sigma})}{\hat{power}(\beta_1, 1)}$$

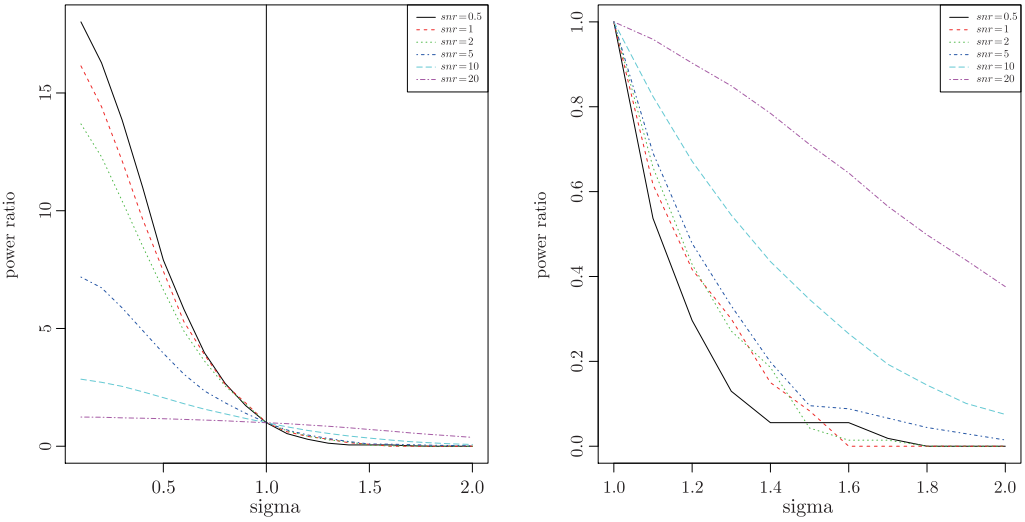

Figure 12. **Left panel:** Ratio of estimated power to estimated power at correct variance estimate ($\hat{\sigma} = 1$) as function of $\hat{\sigma}$ for different values of β_1. Vertical line drawn at $\hat{\sigma} = 1$ for reference. **Right panel:** Same as left panel, only magnifying the area to the right of the vertical line.

at the different settings of β_1 (left) with an magnification of the figure, including only those $\hat{\sigma}$ larger than 1 (right panel). These tell us, for example, that at $\hat{\sigma} = 1.1$ (an upward bias of 10%) and a signal size of $\beta_1 = \sqrt{5}$, $snr = 5$, our power is only 69% of what it would it should be at the correct variance. By the time we get to $\hat{\sigma}^2 = 1.5$, the power is only 9% of what it should be. Recall from Figure 9 that some standard error estimates had median around this value at $snr = 5$. Although not directly comparable with the previous simulation study, this study gives an indication of the dire effects of misestimating the variance.

5. Orthogonal Predictor Matrix and a Certainty Equivalent Variance Estimator

Obtaining finite sample results (or even asymptotic results) about variance estimators with adaptively chosen regularisation parameters seems like a difficult task. In this section, we consider a very simple setup that allows for some tractable results. In particular, since most error variance estimators are based on residual sum of squares, we wish to study the behaviour of a variance estimator based on this quantity in a simple scenario.

Consider the orthogonal case where $p = n$ and $X = I_n$, the $n \times n$ identity matrix. In this case, we have each $Y_i \sim N(\beta_i, \sigma^2)$. We assume that sparsity of the β vector is governed by $\alpha < 1$. In particular, we have $\beta_i = \beta$ for $i = 1, 2, \ldots, \lceil n^\alpha \rceil$ and $\beta_i = 0$ otherwise. Call this the *orthogonal sparsity model*.

Estimates of β_i are obtained by minimising the objective

$$\sum_{i=1}^{n}(Y_i-\beta_i)^2+\lambda\sum_{i=1}^{n}|\beta_i|$$

to obtain the solution $\hat{\beta}_i = S(Y_i,\lambda)$, where $S(x,t)=\text{sign}(x)\max\{|x|-t,0\}$ is the soft thresholding operator with threshold t. Plugging these quantities into the residual sum of squares, we obtain

$$\begin{aligned} RSS &= \sum_{i=1}^{n}(Y_i-\hat{\beta}_i)^2 \\ &= \sum_{i=1}^{n}\min\{Y_i^2,\lambda^2\}, \end{aligned}$$

from which we derive the family of estimators for σ^2 (indexed by λ),

$$\hat{\sigma}^2_{n,\lambda}=\frac{\sum_{i=1}^{n}\min\{Y_i^2,\lambda^2\}}{\sum_{i=1}^{n}I\{|Y_i|\leq\lambda\}}. \tag{5.1}$$

To make this a practicable estimate of σ^2, we need to select a single member from the family (i.e. a value for λ). Many are possible, but in light of the discussion of previous sections, let us select it adaptively. Consider then a single held out set $Z_i \overset{d}{=} Y_i$, independent of Y_i, with corresponding cross-validation error

$$CV(\lambda)=\sum_{i=1}^{n}(Z_i-S(Y_i,\lambda))^2.$$

The adatively chosen regularisation parameter is then

$$\tilde{\lambda}=\text{argmin}_{\lambda}CV(\lambda). \tag{5.2}$$

Its theoretical properties are not considered here, but we believe that $\hat{\sigma}^2_{n,\tilde{\lambda}}$ is amenable to theoretical analysis and that such an analysis may be instructive to the workings of variance estimators with adaptively chosen regularisation parameters. This is definitely a channel for future investigation. However, in the sequel we consider the behaviour of estimators (5.1) under deterministic sequences λ_n. After some general results, we consider a specific sequence, dubbed the *certainty equivalent* (CE) sequence of λ, and denoted $\hat{\lambda}_{n,\alpha,\beta}$, which bears resemblance to the adaptive selection (5.2). Finally, a small simulation reveals how similarly $\hat{\sigma}^2_{n,\tilde{\lambda}}$ and $\hat{\sigma}^2_{n,\hat{\lambda}_{n,\alpha,\beta}}$ behave in small samples, giving hope that the results gleaned for the latter apply to the former.

5.1. General deterministic sequences: λ_n

Our first result considers the large sample behaviour of the denominator in (5.1) under a determinsitic sequence λ_n.

Lemma 1. *If $\lambda_n \to \infty$ as $n \to \infty$, then under the orthogonal sparsity model, $(1/n)\sum_{i=1}^n I\{|Y_i| > \lambda_n\} \xrightarrow{P} 0$.*

All theorems and lemmas are proved in the Appendix. Lemma 1 suggests that we need not consider $\hat{\sigma}^2_{n,\lambda_n}$ directly, but rather the more tractable

$$\tilde{\sigma}^2_{n,\lambda_n} = \frac{1}{n}\sum_{i=1}^n \min\{Y_i^2, \lambda_n^2\}.$$

Lemma 2. *If $\lambda_n \to \infty$ as $n \to \infty$, then under the orthogonal sparsity model $E[\tilde{\sigma}^2_{n,\lambda_n}] \to \sigma^2$ and $\mathrm{Var}\,[\tilde{\sigma}^2_{n,\lambda_n}] \to 0$.*

Theorem 1. *If $\lambda_n \to \infty$ as $n \to \infty$, then under the orthogonal sparsity model*

$$\hat{\sigma}^2_{n,\lambda_n} \xrightarrow{P} \sigma^2,$$
$$\sqrt{n}(\hat{\sigma}^2_{n,\lambda_n} - \sigma^2) \xrightarrow{d} N(0, 2\sigma^4).$$

The consistency and asymptotic normality (with asymptotic variance equal to that of the oracle estimator) are not too hard to come by in this family of estimators. All we need to do is select λ_n that tends to ∞ with n. This is somewhat surprising, but meshes nicely with evidence from our simulation studies. An estimator can have these desirable asymptotic properties, but since the requirement on λ_n to achieve these properties is weak, many consistent, asymptotically normal estimators can have poor finite sample performance.

For example, suppose we set $\lambda_n = \infty$, so that $\hat{\sigma}^2_{n,\lambda_n} = (1/n)\sum_{i=1}^n Y_i$. This estimator satisfies the conditions of Theorem 1, but has finite sample expectation $\sigma^2 + (n^\alpha/n)\beta^2$. We can make this estimator arbitrarily biased upward in a finite sample by merely increasing the signal strength β or reducing sparsity (increasing α). We need a $\lambda_n \to \infty$, but chosen so as to have good finite sample performance as well.

5.2. Certainty equivalent sequence $\hat{\lambda}_{n,\alpha,\beta}$

Instead of choosing λ as the (random) minimiser of $CV(\lambda)$ and inducing dependence between the summands of the numerator of $\hat{\sigma}^2_{n,\tilde{\lambda}}$ (further increasing complexity), we could choose a deterministic sequence (hopefully) bearing close relation to it. In particular, we can minimise $ECV_n(\lambda, \beta, \alpha) = E[CV(\lambda)]$, which can be written as

$$ECV_n(\lambda, \beta, \alpha) = n\sigma^2 + \frac{n^\alpha}{n} \cdot r_S(\lambda, \beta) + \frac{n - n^\alpha}{n} \cdot r_S(\lambda, 0),$$

where $r_S(\lambda, \beta) = E(S(Y_i, \lambda) - \beta)^2$ is the risk of the soft thresholding operator, for which we have the expression (Johnstone (2013))

$$r_S(\lambda, \beta) = \sigma^2 + \lambda^2 + (\beta^2 - \lambda^2 - \sigma^2)\Big[\Phi\Big(\frac{\lambda - \beta}{\sigma}\Big) - \Phi\Big(\frac{\lambda - \beta}{\sigma}\Big)\Big]$$
$$-\sigma(\lambda - \beta)\phi(\lambda + \beta) - \sigma(\lambda + \beta)\phi(\lambda - \beta).$$

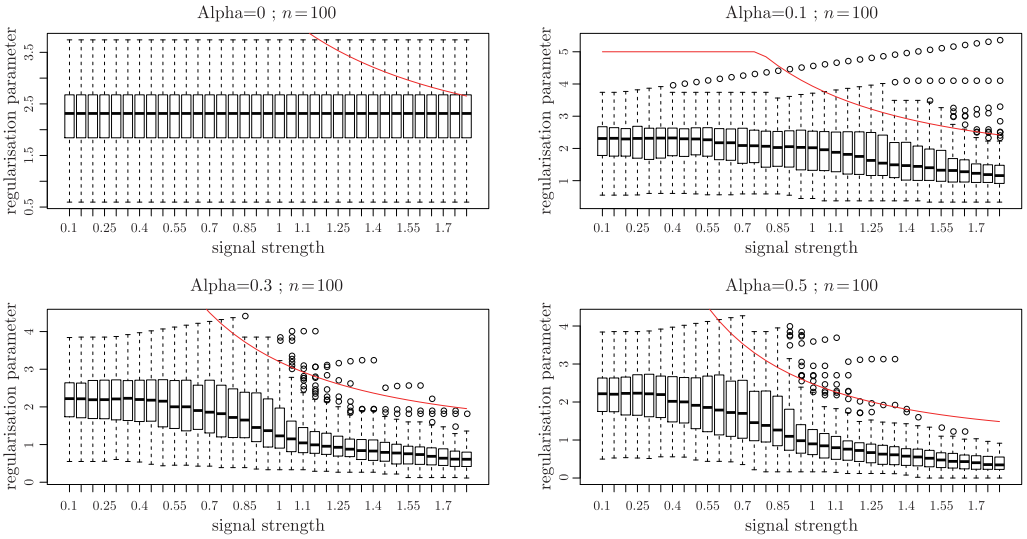

Figure 13. Certainty equivalent regularisation parameter as a function of signal strength (β), at different sparsity levels (α) along with reference boxplots of $CV(\lambda)$ minimising regularisation parameter. Top left panel corresponds to $\alpha = 0$; top right, $\alpha = 0.1$; bottom left, $\alpha = 0.3$ and bottom right, $\alpha = 0.5$. A sample size of $n = 100$ was used when generating replications of CV minimising $\tilde{\lambda}$.

The "certainty equivalent" choice for λ then becomes

$$\hat{\lambda}_{n,\alpha,\beta} = \hat{\lambda}_n(\beta, \alpha) = \text{argmin}_\lambda R_n(\lambda, \beta, \alpha),$$

where $R_n(\lambda, \beta, \alpha) = (n^\alpha/n) r_S(\lambda, \beta) + [(n - n^\alpha)/n] r_S(\lambda, 0)$.

The optimisation can be done numerically. Figure 13 plots $\hat{\lambda}_{n,\alpha,\beta}(\beta, \alpha)$ as a function of the signal strength β for four different levels of sparsity α (solid lines). Also plotted in Figure 13 are boxplots gleaned from $B = 100$ realisations of $\tilde{\lambda}$ where

$$\tilde{\lambda} = \text{argmin}_\lambda CV(\lambda)$$

for a sample of size $n = 100$. We plot these for reference, because they are the realisations of the random regularisation parameter we would actually compute in an application. Notice how $\hat{\lambda}_{n,\alpha,\beta}$ tends to lie everywhere above the rump of the $\tilde{\lambda}$ values at a given signal strength, exhibiting a similar shape. Although convenient theoretically, the certainty equivalent estimate of λ does not seem to accord with the random estimate obtained by minimising $CV(\lambda)$. Despite this, their estimates for σ^2 are not too different, as demonstrated in the next section.

This sequence cannot be obtained in applications, because we do not know σ^2 nor β. The certainty equivalent sequence can only be generated by an oracle.

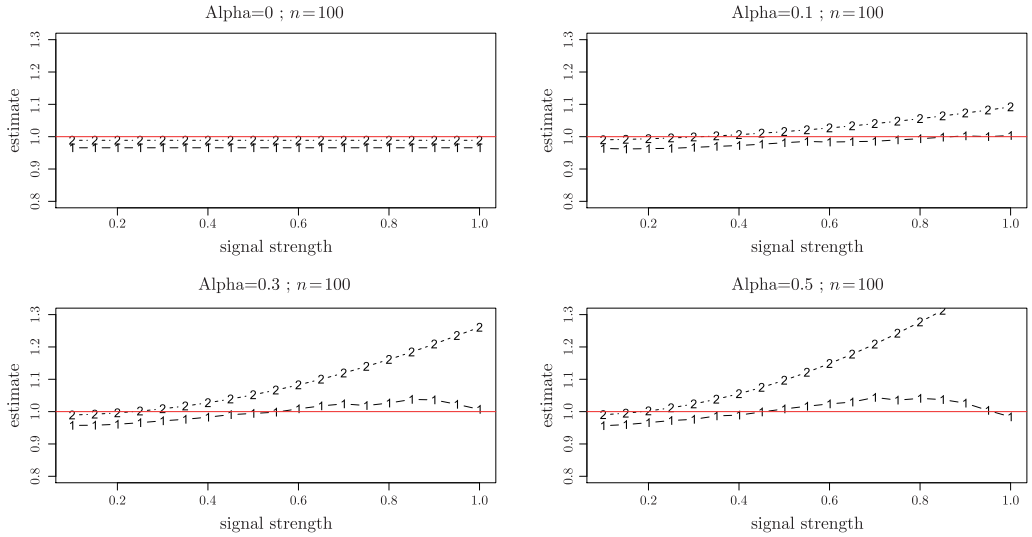

Figure 14. Variance estimates as a function of signal strength (β), at different sparsity levels (α). Curves labelled "1" plot the CV estimate, while those labelled "2" and "3" the CE estimates (with different denominators). Sparsity parameter $\alpha = 0$, 0.1, 0.3 and 0.5, varying left to right along rows and then down along columns. Horizontal lines show the true variance.

The true utility of this sequence comes from its theoretical tractability. As a function of λ, $r_S(\lambda, 0)$ is monotonically decreasing, convex, and non-negative. Its minimum value is 0, achieved by setting $\lambda = \infty$. For $\beta \neq 0$, $r_S(\lambda, \beta)$ is initially decreasing in λ, attains a unique global minimum, whereafter is non-decreasing in λ, with horizontal asymptote β^2. Similar properties are shared by $(n^\alpha/n)r_S(\beta, \lambda) + [(n - n^\alpha)/n]r_S(\lambda, 0)$.

We can show that $\hat{\lambda}_{n,\alpha,\beta} \to \infty$, making $\hat{\sigma}^2_{n,\hat{\lambda}_{n,\alpha,\beta}}$ consistent for σ^2 and ensuring it has an asymptotic normal distribution (from the previous section). Furthermore, one can show how this sequence minimises an upper bound to the upward bias of the estimator in small samples. Downward bias does not seem to be a problem for this estimator (see Figure 14).

Theorem 2. *Assume the orthogonal sparsity model. The certainty equivalent sequence* $\hat{\lambda}_{n,\alpha,\beta} \to \infty$ *as* $n \to \infty$.

To see how this sequence minimises an upper bound on the upward bias, consider the Stein Unbiased Risk Estimate (SURE) for $r_S(\lambda, \beta)$ (Johnstone (2013)):

$$r_S(\lambda, \beta) = E[\sigma^2 - 2\sigma^2 I\{|Y| \leq \lambda\} + \min\{Y^2, \lambda^2\}],$$

where $Y \sim N(\beta, \sigma^2)$, so that

$$E[\min\{Y_i^2, \lambda\}] - \sigma^2 = r_S(\lambda, \beta_i) - 2P(|Y_i| > \lambda)$$

$$\leq r_S(\lambda, \beta_i),$$

which translates to

$$E[\tilde{\sigma}^2_{n,\lambda_n}] - \sigma^2 \leq \frac{n^\alpha}{n} r_S(\lambda_n, \beta) + \frac{n - n^\alpha}{n} r_S(\lambda_n, 0),$$

the right hand side of which is minimised by the certainty equivalent sequence of λ. The certainty equivalent sequence minimises an upper bound to the bias, making a concerted effort to keep it as small as possible *in the finite sample*.

5.3. Comparison of CE and CV variance estimators

Figure 14 plots, for different levels of sparsity in different panels, the mean over $B = 100$ replications of three variance estimators. The curves labelled "1" are the means of the replications of $\hat{\sigma}^2_{CV} = \hat{\sigma}^2_{n,\tilde{\lambda}}$, where $\tilde{\lambda}$ is chosen according to (5.2). Curves labelled "2" are of the means of the replications of $\hat{\sigma}^2_{CE} = \hat{\sigma}^2_{n,\hat{\lambda}_{n,\alpha,\beta}}$, where $\hat{\lambda}_{n,\alpha,\beta}$ is the certainty equivalent sequence of λ, while those labelled "3" are of $\tilde{\sigma}^2_{CE} = \tilde{\sigma}^2_{n,\hat{\lambda}_{n,\alpha,\beta}}$. Notice how close curves "2" and "3" are to each other. We have the theoretical guarantee on the bias of curves "3".

The three estimators behave reasonably similarly, except for low sparsity, high signal cases. It is heartening to note that the CE estimate seems to suffer from upward bias in this case, despite its guarantee of a minimum upper bound on this bias. The CV estimator actually seems to do an even better job of selecting the appropriate regularisation parameter in the small sample setting - a clear case for further analysis of its properties.

6. Discussion

Error variance estimation in linear regression when $p > n$ is a difficult problem that deserves attention. Several estimators have been proposed. We have reviewed these and some of the theoretical results around them. Despite some comforting asymptotic results, finite sample performance of these estimators seems to suffer, particularly when signals become large and non-sparse.

Variance estimators based on residual sums of squares with adaptively chosen regularisation parameters seem to have promising finite sample properties. In particular, we recommend the cross-validation based, Lasso residual sum of squares estimator as a good variance estimator under a broad range of sparsity and signal strength assumptions. The complexity of their structure seem to discourage their rigorous analysis. Simulation results from this paper suggest that there could be value in understanding these estimators more fully.

Acknowledgements

We wish to thank Iain Johnstone for his helpful comments that seeded the analysis of the certainty equivalent in the orthogonal case.

We also wish to thank two anonymous referees for their helpful and insightful comments. The changes suggested by them have truly improved the quality of the paper.

Appendix. Proofs of lemmas and theorems

Proof of Lemma 1. Assume the orthogonal sparsity model and consider

$$\begin{aligned}
&E\left[\frac{1}{n}\sum_{i=1}^{n} I\{|Y_i| > \lambda_n\}\right] = \frac{1}{n}\sum_{i=1}^{n} P(|Y_i| > \lambda_n) \\
&= \frac{n^\alpha}{n}\left(1 - \Phi\left(\frac{\lambda_n - \beta}{\sigma}\right) + \Phi\left(\frac{-\lambda_n - \beta}{\sigma}\right)\right) + 2\frac{n - n^\alpha}{n}\left(1 - \Phi\left(\frac{\lambda_n}{\sigma}\right)\right) \\
&\leq 2\frac{n^\alpha}{n} + 2\frac{n - n^\alpha}{n}\left(1 - \Phi\left(\frac{\lambda_n}{\sigma}\right)\right) \\
&\to 0
\end{aligned}$$

as $n, \lambda_n \to \infty$. The result follows upon the application of Markov's inequality.

Proof of Lemma 2. Assume the orthogonal sparsity model and $\lambda_n \to \infty$ as $n \to \infty$ and consider:

$$\begin{aligned}
&E\left[\min\{Y_i^2, \lambda_n^2\}\right] \\
&\quad = E[Y_i^2; -\lambda_n \leq Y_i \leq \lambda_n] + \lambda_n^2 P(|Y_i| > \lambda_n) \\
&\quad = E[Y_i^2; -\lambda_n \leq Y_i \leq \lambda_n] + \lambda_n^2\left(1 - \Phi\left(\frac{\lambda_n - \beta_i}{\sigma}\right)\right) + \lambda_n^2\left(1 - \Phi\left(\frac{\lambda_n + \beta_i}{\sigma}\right)\right).
\end{aligned}$$

Now

$$E[Y_i^2; -\lambda_n \leq Y_i \leq \lambda_n] = \sigma^2 m_2(\lambda_n, \beta_i) + 2\sigma\beta_i m_1(\lambda_n, \beta_i) + \beta_i^2 m_0(\lambda_n, \beta_i),$$

where

$$m_j(\lambda, \beta) = \int_{\frac{-\lambda-\beta}{\sigma}}^{\frac{\lambda-\beta}{\sigma}} x^j \phi(x)\, dx$$

for all non-neative integers j. Note that for fixed β and σ and $\lambda \to \infty$, these tend to the j^{th} moments of the standard normal distribution. So $m_0(\lambda_n, \beta_i) \to 1$, $m_1(\lambda_n, \beta_i) \to 0$ and $m_2(\lambda_n, \beta_i) \to 1$ as $n \to \infty$. This suggests that $E[Y_i^2; -\lambda_n \leq Y_i \leq \lambda_n] \to \sigma^2 + \beta_i^2$ as $n \to \infty$.

Also note that for $x > 0$, as $x \to \infty$, $x^k\,(1 - \Phi(x)) \sim x^{k-1}\phi(x) \to 0$ for any finite integer $k > 1$. Hence

$$\begin{aligned} E[\tilde{\sigma}^2_{n,\lambda_n}] &\sim \frac{n^\alpha}{n}(\sigma^2+\beta^2)+\frac{n-n^\alpha}{n}\sigma^2 \\ &\to \sigma^2 \end{aligned}$$

as $n \to \infty$.

Similarly,

$$\begin{aligned} &E[\min\{Y_i^4, \lambda^4\}] \\ &= E[Y_i^4; -\lambda_n \le Y_i \le \lambda_n] + \lambda_n^4\left(1-\Phi\left(\frac{\lambda_n-\beta_i}{\sigma}\right)\right)+\lambda_n^4\left(1-\Phi\left(\frac{\lambda_n+\beta_i}{\sigma}\right)\right) \end{aligned}$$

with

$$\begin{aligned} E[Y_i^4; -\lambda_n \le Y_i \le \lambda_n] &= \sigma^4 m_4(\lambda_n,\beta_i)+4\sigma^3\beta_i m_3(\lambda_n,\beta_i)+6\sigma^2\beta_i^2 m_2(\lambda_n,\beta_i) \\ &\quad + 4\sigma\beta_i^3 m_1(\lambda_n,\beta_i)+\beta_i^4 m_0(\lambda_n,\beta_i) \\ &\sim 3\sigma^4+6\sigma^2\beta_i^2+\beta_i^4. \end{aligned}$$

Also

$$\begin{aligned} E[\min\{Y_i^2,\lambda_n^2\}\min\{Y_j^2,\lambda_n^2\}] &= E[\min\{Y_i^2,\lambda_n^2\}]E[\min\{Y_j^2,\lambda_n^2\}] \\ &\sim \sigma^4+\sigma^2\beta_i^2+\sigma^2\beta_j^2+\beta_i^2\beta_j^2. \end{aligned}$$

So that

$$\begin{aligned} E[\tilde{\sigma}^4_{n,\lambda_n}] &\sim \frac{n^\alpha}{n^2}(3\sigma^4+6\sigma^2\beta^2+\beta^4)+\frac{n-n^\alpha}{n^2}(3\sigma^4)+\frac{n^\alpha(n^\alpha-1)}{n^2}(\sigma^4+2\sigma^2\beta^2+\beta^4) \\ &\quad + \frac{n^\alpha(n-n^\alpha)}{n^2}(\sigma^4+\sigma^2\beta^2)+\frac{(n-n^\alpha)(n-n^\alpha-1)}{n^2}(\sigma^4) \\ &\to \sigma^4 \end{aligned}$$

as $n \to \infty$. Hence $Var[\tilde{\sigma}^2_{n,\lambda_n}] \to 0$.

Proof of Theorem 1. An application of Markov's inequality to $(\tilde{\sigma}^2_{n,\lambda_n}-\sigma^2)^2$, combined with the results of Lemma 2 give us consistency of $\tilde{\sigma}^2_{n,\lambda_n}$ for σ^2. Combined with the result of Lemma 1, we have that $\hat{\sigma}^2_{n,\lambda_n}$ is consistent for σ^2.

Asymptotic normality of $\tilde{\sigma}^2_{n,\lambda_n}$ follows from a central limit theorem applied to the *independent* summands of the numerator. The asymptotic variance is gleaned from the proof of Lemma 2, by noting that the results quoted there imply that $nVar[\tilde{\sigma}^4_{n,\lambda_n}] \sim 3\sigma^4-\sigma^4 = 2\sigma^4$. Lemma 1 ensures that this asymptotic normality holds for $\hat{\sigma}^2_{n,\lambda}$ as well.

Proof of Theorem 2. Fix α and β and let $Q_n(\lambda) = R_n(\lambda,\beta,\alpha)$ and $Q_0(\lambda) = r_S(\lambda,0)$.

Now

$$\begin{aligned}
|Q_n(\lambda) - Q_0(\lambda)| &= \frac{n^\alpha}{n}|r_S(\lambda,\beta) - r_S(\lambda,0)| \\
&\leq \frac{n^\alpha}{n}\left[r_S(\lambda,\beta) + r_S(\lambda,0)\right] \\
&\leq \frac{n^\alpha}{n}\left[2r_S(\lambda,0) + \beta^2\right] \\
&\leq \frac{n^\alpha}{n}\left[2\sigma^2 + \beta^2\right] \\
&\to 0
\end{aligned}$$

as $n \to \infty$, uniformly in λ. The third and fourth lines follow from Johnstone (2013).

For any $\epsilon > 0$, we have from the fact that $\hat{\lambda}_{n,\alpha,\beta} = \text{argmin}_{\lambda \geq 0} Q_n(\lambda)$ and $(n^\alpha/n)\beta^2 \to 0$, that $\exists N_1 = N_1(\epsilon)$ such that $\forall n > N_1$,

$$\begin{aligned}
Q_n(\hat{\lambda}_{n,\alpha,\beta}) &< Q_n(\infty) + \frac{\epsilon}{3} \\
&= \frac{n^\alpha}{n}\beta^2 + \frac{\epsilon}{3} < \frac{2\epsilon}{3}.
\end{aligned}$$

Furthermore, by the uniform convergence proven above, $\exists N_2 = N_2(\epsilon)$ such that $\forall n > N_2$, $Q_0(\hat{\lambda}_{n,\alpha,\beta}) < Q_n(\hat{\lambda}_{n,\alpha,\beta}) + \epsilon/3$. Hence, $\forall n > \max\{N_1, N_2\}$, $Q_0(\hat{\lambda}_{n,\alpha,\beta}) < \epsilon$.

Let $M > 0$. Since the set $[0, M]$ is closed and compact, $Q_0(\lambda^\star_M) = \inf_{\lambda \in [0,M]} Q_0(\lambda)$ attains. Note $Q_0(\lambda^\star_M) > Q_0(\infty) = 0$. Choosing $\epsilon = \epsilon(M) = Q_0(\lambda^\star_M)$, we have $Q_0(\hat{\lambda}_{n,\alpha,\beta}) < Q_0(\lambda^\star_M)$, so that $\hat{\lambda}_{n,\alpha,\beta} > M$. Since M is arbitrary, we have $\hat{\lambda}_{n,\alpha,\beta} \to \infty$.

References

Dicker, L. H. (2014). Variance estimation in high-dimensional linear models. Biometrika, doi: 10.1093/biomet/ast065.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Fan, J., Guo, S. and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. Roy. Statist. Soc. Ser. B* **75**, 37-65.

Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10**, 971-988.

Homrighausen, D. and McDonald, D. J. (2013). The lasso, persistence, and cross-validation. In *Proceedings of the* 30*th International Conference on Machine Learning*.

Javanmard, A. and Montanari, A. (2013). Confidence intervals and hypothesis testing for high-dimensional regression. arXiv:1306.3171.

Johnstone, I. (2013). Gaussian estimation: Sequence and wavelet models. Stanford University.

Lockhart, R., Taylor, J., Tibshirani, R. and R. Tibshirani, R. (2013). A significance test for the lasso. Stanford University, arXiv:1301.7161.

Stadler, N. and Buhlmann, P. and van der Geer, S. (2010). ℓ_1-penalization for mixture regression models. *Test* **19**, 206-256.

Sun, T. and Zhang, C.-H. (2010). Comments on: ℓ_1-penalization for mixture regression models. *Test* **19**, 270-275.

Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99**, 879-898.

Sun, T. and Zhang, C.-H. (2013). Sparse matrix inversion with scaled lasso. *J. Machine Learn. Res.* **14**, 3385-3418.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Wasserman, L. and Roeder, K. (2009). High-dimensional variable selection. *Ann. Statist.* **37**, 2178-2201.

Yu, Y. and Feng, Y. (2013). Modified cross-validation for penalized high-dimensional linear regression models. *J. Comput. Graph. Statist.* **23**, 1009-1027.

Department of Statistics, Stanford University, Stanford, CA 94305, USA.

E-mail: sreid@stanford.edu

Departments of Health, Research & Policy, and Statistics, Stanford University, Stanford, CA 94305, USA.

E-mail: tibs@stanford.edu

Department of Statistics, Stanford University, Stanford, CA 94305, USA.

E-mail: jhf@stanford.edu

(Received February 2014; accepted December 2014)