On the Use of Information Criteria for Subset Selection in Least Squares Regression

Sen Tian* Clifford M. Hurvich Jeffrey S. Simonoff

Department of Technology, Operations, and Statistics,
Stern School of Business, New York University.

Abstract

Least squares (LS) based subset selection methods are popular in linear regression modeling when the number of predictors is less than the number of observations. Best subset selection (BS) is known to be NP hard and has a computational cost that grows exponentially with the number of predictors. Forward stepwise selection (FS) is a greedy heuristic for BS. Both methods rely on cross-validation (CV) in order to select the subset size k, which requires fitting the procedures multiple times and results in a selected k that is random across replications. Compared to CV, information criteria only require fitting the procedures once, and we show that for LS-based methods they can result in better predictive performance while providing a non-random choice of k. However, information criteria require knowledge of the effective degrees of freedom for the fitting procedure, which is generally not available analytically for complex methods. In this paper, we propose a novel LS-based method, the best orthogonalized subset selection (BOSS) method, which performs BS upon an orthogonalized basis of ordered predictors. Assuming orthogonal predictors, we build a connection between BS and its Lagrangian formulation (i.e., minimization of the residual sum of squares plus the product of a regularization parameter and k), and based on this connection introduce a heuristic degrees of freedom (hdf) for BOSS that can be estimated via an analytically-based expression. We show in both simulations and real data analysis that BOSS using the Kullback-Leibler based information criterion AICc-hdf has the strongest performance of all of the LS-based methods considered and is competitive with regularization methods, with the computational effort of a single ordinary LS fit. Supplementary materials are attached at the end of the main document. An R package **BOSSreg** and the computer code to reproduce the results for this article are available online[1].

Keywords: Least squares; Best subset selection; Effective degrees of freedom; Information criteria; Cross validation.

1 Introduction

Suppose that we have the data generating process

$$\mathbf{y} = \mu + \epsilon, \tag{1}$$

where $\mathbf{y} \in \mathcal{R}^n$ is the response vector, $\mu \in \mathcal{R}^n$, is the fixed mean vector, and $\epsilon \in \mathcal{R}^n$ is the noise vector. The mean vector is estimated based on a fixed design matrix $\mathbf{X} \in \mathcal{R}^{n \times p}$. We assume the error $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ and $n > p$.

*E-mail: stian@stern.nyu.edu

[1] `https://github.com/sentian/BOSSreg`. At the time of submission of this paper, we have submitted the R package to *CRAN*.

1.1 Best subset selection

Best subset selection (BS) (Hocking and Leslie, 1967) seeks the set of predictors that best fit the data in terms of quadratic error for each given subset size (excluding the intercept) $k \in \{0, 1, \cdots, p\}$, i.e. it solves the following constrained optimization problem:

$$\min_{\beta_0,\beta} \frac{1}{2}\|\mathbf{y} - \beta_0 \mathbf{1} - \mathbf{X}\beta\|_2^2 \quad \text{subject to} \quad \|\beta\|_0 \leq k, \tag{2}$$

where $\|\beta\|_0 = \sum_{i=1}^{p} \mathbf{1}(\beta_i \neq 0)$ is the number of non-zero coefficients in β. Note that to simplify the discussion, we assume that the intercept term $\beta_0 = 0$ throughout the paper, except in the real data examples, where all of the fitting procedures include an intercept.

BS is known to be an NP-hard problem (Natarajan, 1995) and its computational cost grows exponentially with the dimension p. Many attempts have been made to reduce the computational cost of the method. The most well-known approach is the branch-and-bound algorithm 'leaps' (Furnival and Wilson, 1974) that solves (2) in seconds for p being up to around 30. More recently, Bertsimas et al. (2016) formulated (2) using a mixed integer operator (MIO), and largely reduced the computing overhead by using a well-developed optimization solver such as `GUROBI` or `CPLEX`. However, according to Hastie et al. (2017), MIO normally takes about 3 minutes to find the solution at a given size k for a problem with $n = 500$ and $p = 100$. The current methodology is not scalable to very large datasets, and solving (2) remains a challenge for most real world applications.

In order to select the optimal tuning parameter, e.g. the subset size k in (2), one often applies an information criterion, which augments the training error with the effective degrees of freedom (edf). Efron (1986) defined the edf for a general fitting rule $\hat{\mu}$ as:

$$\text{edf}(\hat{\mu}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \text{cov}(\hat{\mu}_i, y_i). \tag{3}$$

It is easy to verify that edf for the linear regression of y upon X is the number of estimated coefficients p. However, Janson et al. (2015) showed in simulations that there can be a large discrepancy between the edf of BS at size k and k itself. Similar evidence can be found in Tibshirani (2015), where the author quantifies the difference as the search degrees of freedom, which accommodates the amount of searching that BS performs in order to choose a best k-predictor subset. Unfortunately, the edf of BS does not have an analytical expression except when X is orthogonal and $\mu = 0$ (Ye, 1998). Numerically, we can apply tools like data perturbation (Ye, 1998), bootstrap (Efron, 2004) or data randomization (Harris, 2016) to estimate edf, but all rely on tunings of some hyperparameters and can be computationally intensive.

This paper is motivated by the above challenges. We propose a heuristic degrees of freedom (hdf) for BS by assuming an orthogonal X. We further introduce a novel least squares (LS) based subset selection method, the best orthogonalized subset selection (BOSS), and we demonstrate that BOSS using an information criterion with hdf works well in practice for a general X with the computational cost of a single ordinary LS fit.

1.2 Optimism theorem and information criteria for BS

Information criteria are designed to provide an unbiased estimate of the prediction (or testing) error, and can be derived from the so-called optimism theorem. Denote Θ as an error measure, err as the training error, Err as the testing error, y^0 as a new response vector with the same distribution but independent of the original y, and E_0 is the expectation taken over y^0. Efron (1986) defined the optimism as

$$\text{op} = \text{Err} - \text{err},$$

and introduced the optimism theorem,

$$E(\text{op}) = E(\text{Err}) - E(\text{err}).$$

A straightforward result from the optimism theorem is that

$$\widehat{\text{Err}} = \text{err} + E(\text{op}) \tag{4}$$

is an unbiased estimator of $E(\text{Err})$, and is intended to balance the trade-off between model fit and model complexity. The challenge is to find $E(\text{op})$ for a given fitting rule $\hat{\mu}$ and error measure $\Theta(y, \hat{\mu})$.

When the error measure Θ is the squared error (SE), i.e. $\Theta(y_i, \hat{\mu}_i) = (y_i - \hat{\mu}_i)^2$, err_{SE} (denoted as the training error when Θ is SE) then becomes the residual sum of squares $\text{RSS} = \sum_{i=1}^{n} \Theta(y_i, \hat{\mu}_i)$, and the testing error $\text{Err}_{\text{SE}} = \sum_{i=1}^{n} E_0[\Theta(y_i^0, \hat{\mu}_i)]$. Ye (1998) and Efron (2004) proved that for a general fitting rule $\hat{\mu}$ such as BS, $E(\text{op}_{\text{SE}}) = 2\sigma^2 \cdot \text{edf}(\hat{\mu})$, and hence $\widehat{\text{Err}}_{\text{SE}}$ in (4) becomes C_p-edf where

$$\text{C}_p\text{-edf} = \text{RSS} + 2\sigma^2 \cdot \text{edf}. \tag{5}$$

These authors also showed that the traditional C_p,

$$\text{C}_p\text{-ndf} = \text{RSS} + 2\sigma^2 \cdot \text{ndf},$$

can be greatly biased when applied for BS, since C_p-ndf (Mallows, 1973) was derived for a linear estimation rule $\hat{\mu} = Hy$ where H is independent of y, which is not the case for BS. Here ndf denotes the naive degrees of freedom, i.e. $\text{Tr}(H)$. A further major issue regarding applying C_p in practice is that it requires an estimate of σ^2.

Another commonly used error measure is the deviance (up to a constant)

$$\Theta = -2 \log f(y|\mu, \sigma^2), \tag{6}$$

where f is a pre-specified parametric model. Let $\hat{\mu}$ and $\hat{\sigma}^2$ be the maximum likelihood estimators obtained by maximizing $f(y|\mu, \sigma^2)$. We then have $\text{err}_{\text{KL}} = -2 \log f(y|\hat{\mu}, \hat{\sigma}^2)$ and $\text{Err}_{\text{KL}} = -2E_0\left[\log f(y^0|\hat{\mu}, \hat{\sigma}^2)\right]$, where the latter is the Kullback-Leibler (KL) discrepancy. For a linear estimation procedure, assuming asymptotic normality of $\hat{\mu}$ and $\hat{\sigma}^2$ (f not necessarily Gaussian) and the true model distribution being contained in the specified parametric model f, Konishi and Kitagawa (2008) proved that $E(\text{op}_{\text{KL}}) = 2 \cdot \text{ndf} + o(1)$, and AIC (Akaike, 1973),

$$-2 \log f(y|\hat{\mu}, \hat{\sigma}^2) + 2 \cdot \text{ndf},$$

asymptotically equals $\widehat{\text{Err}}_{\text{KL}}$ (4). If f follows a Gaussian distribution, as assumed in (1), AIC can be expressed as

$$\text{AIC-ndf} = n \log\left(\frac{\text{RSS}}{n}\right) + 2 \cdot \text{ndf}.$$

Hurvich and Tsai (1989) replaced the asymptotic $E(\text{op}_{\text{KL}})$ with its exact value, for Gaussian linear regression with an assumption that the predictors with non-zero true coefficients are included in the model, and used the corrected AIC

$$\text{AICc-ndf} = n \log\left(\frac{\text{RSS}}{n}\right) + n\frac{n + \text{ndf}}{n - \text{ndf} - 2}.$$

Neither AIC nor AICc has a penalty term depending upon σ^2, a clear advantage over C_p.

It remains a challenge to derive a KL-based information criterion for BS. Liao et al. (2018) estimated $E(\text{op}_{\text{KL}})$ via Monte Carlo simulations, but this relies on thousands of fits of the procedure, which is not computationally feasible for large datasets.

In this work, we propose AICc-edf

$$\text{AICc-edf} = n \log\left(\frac{\text{RSS}}{n}\right) + n\frac{n + \text{edf}}{n - \text{edf} - 2} \tag{7}$$

for this purpose. We demonstrate that $E(\text{AICc-edf})$ approximates $E(\text{Err}_{\text{KL}})$ well for BS. Moreover, both AICc-edf and $\widehat{\text{Err}}_{\text{KL}}$ generally choose the same subset when used as selection rules. Furthermore, the feasible implementation AICc-hdf works reasonably well as a selection rule for BS with an orthogonal X and for our proposed method BOSS with a general X.

1.3 The structure and contributions of this paper

The rest of the paper is organized as follows. In Section 2, by assuming an orthogonal X, we introduce the hdf for BS. We provide a theoretical justification in a restricted scenario, and numerical justifications in general situations. We provide numerical evidence that $E(\text{AICc-edf})$ approximates $E(\text{Err}_{\text{KL}})$ well, and the feasible version AICc-hdf works well as a selection rule for BS. We further compare the performance of BS with that of regularization methods using feasible selection rules via simulations. In Section 3, we consider a general X and propose the method BOSS. We provide numerical evidence that AICc-hdf is a reasonable selection rule for BOSS. Furthermore, we compare the performance of BOSS with that of forward stepwise regression (FS) and regularization methods in simulations. Lastly, we study some real data examples in Section 4, and provide conclusions and potential future works in Section 5.

Below is guidance for applying LS-based methods in practice for data analysts.

- Using information criteria in a naive way by plugging in the subset size as the degrees of freedom can lead to significantly worse performance than using edf and the feasible hdf.
- AICc is a better selection rule in terms of predictive performance in comparison to C_p, and the advantage is particularly strong when p is close to n.
- AICc is not only more computationally efficient than cross-validation (CV), but also can result in subsets with better predictive performance especially when the signal-to-noise ratio (SNR) is high or the sample size n is large. The SNR is defined as $\text{Var}(x^T\beta)/\sigma^2$.
- BOSS-AICc is generally the best LS-based method in comparison to BS and forward stepwise (FS) (Efroymson, 1960) using CV as the selection rule, in terms of both computational efficiency and predictive performance.
- Compared to regularization methods, BOSS-AICc performs the best when SNR is high or the sample size n is large. In terms of support recovery in a sparse true model, BOSS recovers the true predictors and rarely includes any false positives when SNR is high or the sample size n is large. In contrast, regularization methods generally overfit.

2 AICc-hdf for BS with orthogonal X

2.1 A heuristic degrees of freedom for BS

The edf of BS has an analytical expression only when the true model is $\mu = 0$ (Ye, 1998). Tibshirani (2015) studied the Lagrangian formulation of BS (LBS) and provided an analytical expression for edf without any restrictions on μ. To distinguish between the two methods, we use $\text{df}_C(k)$ and $\text{df}_L(\lambda)$ to denote edf of BS for subset size k and edf of LBS for tuning parameter λ, respectively. In this section, we introduce a heuristic degrees of freedom (hdf) for BS that is built upon the connection between $\text{df}_C(k)$ and $\text{df}_L(\lambda)$.

2.1.1 Lagrangian BS and its edf

For each regularization parameter $\lambda \geq 0$, LBS solves

$$\min_{\beta} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_0. \tag{8}$$

Both LBS (8) and BS (2) are LS regressions of y upon a certain subset of X. With orthogonal X, both problems have analytical solutions: $\hat{\beta}_i(\lambda) = z_i \mathbb{1}_{(|z_i| \geq \sqrt{2\lambda})}$ for (8) and $\hat{\beta}_i(k) = z_i \mathbb{1}_{(|z_i| \geq |z_{(k)}|)}$ for (2), where $z = X^T y$ and $z_{(k)}$ is the k-th largest coefficient in absolute value. These two problems are not equivalent, and there is no clear one-to-one correspondence between λ in (8) and k in (2). Indeed, for each λ there exists a k such that $\hat{\beta}(\lambda) = \hat{\beta}(k)$ where $\hat{\beta}(\lambda)$ is the solution of (8) at λ and $\hat{\beta}(k)$ is the solution of (2) at k, but the reverse does not necessarily hold, since there will be multiple λ corresponding to the same solution $\hat{\beta}(k)$. Moreover, with a general X, solving (8) does not guarantee recovery of the entire solution path given by solving (2) for $k = 0, \ldots, p$.

By assuming an orthogonal X, Tibshirani (2015) derived an expression for $\mathrm{df}_L(\lambda)$ based on definition (3),

$$\mathrm{df}_L(\lambda) = E(k_L(\lambda)) + \frac{\sqrt{2\lambda}}{\sigma} \sum_{i=1}^{p} \left[\phi\left(\frac{\sqrt{2\lambda} - (X^T\mu)_i}{\sigma} \right) + \phi\left(\frac{-\sqrt{2\lambda} - (X^T\mu)_i}{\sigma} \right) \right], \tag{9}$$

where the expected subset size is given as

$$E(k_L(\lambda)) = \sum_{i=1}^{p} \left[1 - \Phi\left(\frac{\sqrt{2\lambda} - (X^T\mu)_i}{\sigma} \right) + \Phi\left(\frac{-\sqrt{2\lambda} - (X^T\mu)_i}{\sigma} \right) \right]. \tag{10}$$

2.1.2 hdf for BS

Given the similarity of problems (2) and (8), we would like to approximate $\mathrm{df}_C(k)$ with $\mathrm{df}_L(\lambda)$. One implementation of this proceeds as follows. Note that $\mathrm{df}_C(k)$ is a discrete function of $k = 0, \cdots, p$ while $\mathrm{df}_L(\lambda)$ is a continuous function of a real variable $\lambda \geq 0$. We propose an hdf that uses $\mathrm{df}_L(\lambda)$ for a particular value of λ depending on k as a proxy for $\mathrm{df}_C(k)$. Based on (10), λ and $E(k_L(\lambda))$ have a clear one-to-one correspondence, which implies that we can find a unique $\lambda_k^\star$ such that $E(k_L(\lambda_k^\star)) = k$ for each $k = 1, \cdots, p$. The value of hdf is $\mathrm{df}_L(\lambda_k^\star)$ obtained by substituting $\lambda_k^\star$ into (9). We also let $\mathrm{hdf}(0) = 0$ since $\mathrm{df}_C(0) = 0$. The implementation process is summarized in Algorithm 1. Note that hdf requires estimates of μ and σ, and we use the estimates from the LS regression on all predictors.

Algorithm 1 The heuristic df (hdf) of BS for size k

Input: X (orthogonal), σ and μ. For a given subset size k,

1. Based on (10), calculate $\lambda_k^\star$ such that $E(k_L(\lambda_k^\star)) = k$.
2. Based on (9), calculate $\mathrm{hdf}(k) = \mathrm{df}_L(\lambda_k^\star)$.

Repeat the above steps for $k = 1, \cdots, p$ and let $\mathrm{hdf}(0) = 0$, yielding hdf for each subset.
In place of μ and σ in (9) and (10), we use OLS estimates based on the full model, i.e. $\hat{\mu} = XX^T y$, $\hat{\sigma}^2 = \|y - \hat{\mu}\|_2^2/(n-p)$.

2.1.3 Theoretical justification of hdf under a null true model

Assume $\mu = 0$, with X still being orthogonal. In such a restricted scenario, $\mathrm{df}_C(k)$ has an analytical expression, which allows us to provide some theoretical justification for $\mathrm{hdf}(k)$. We start by introducing

notation, and present the main result in Theorem 1 and its Corollary. The detailed proofs are given in the Supplemental Material.

Denote $\tilde{X}_{(i)}$ as the i-th largest order statistic in an i.i.d sample of size p from a χ^2_1 distribution. Ye (1998) showed that

$$\mathrm{df}_C(k) = E\left(\sum_{i=1}^{k} \tilde{X}_{(i)}\right).$$

Let $\tilde{H}(s) = -\tilde{Q}(1-s)$ where $\tilde{Q}$ is the quantile function of a χ^2_1 distribution, and $s \in (0,1)$. For $0 \le s \le t \le 1$, the truncated variance function is defined as

$$\tilde{\sigma}^2(s,t) = \int_s^t \int_s^t (u \wedge v - uv) d\tilde{H}(u) d\tilde{H}(v),$$

where $u \wedge v = \min(u,v)$. Denote $\tilde{Y}_p = \tilde{\sigma}_p^{-1}(\sum_{i=1}^{k} \tilde{X}_{(i)} - \tilde{\mu}_p)$, where

$$\tilde{\sigma}_p = \sqrt{p} \cdot \tilde{\sigma}(1/p, k/p),$$

and

$$\tilde{\mu}_p = -p \int_{1/p}^{k/p} \tilde{H}(u) du - \tilde{H}\left(\frac{1}{p}\right).$$

Theorem 1. *Assume X is orthogonal and the true model is null ($\mu = 0$). As $p \to \infty$, $k \to \infty$ with $k = \lfloor px \rfloor$, we have*

$$\frac{1}{2p} hdf(k) = \frac{1}{2p} df_C(k) - \frac{\tilde{\sigma}_p}{2p} E(\tilde{Y}_p) + O\left(\frac{\log(p)}{p}\right), \tag{11}$$

where $x \in (0,1)$ is a constant and $\lfloor \cdot \rfloor$ denotes the greatest integer function.

Corollary 1.1. *If $\limsup |E(\tilde{Y}_p)| < \infty$, we further have*

$$\frac{df_C(k)}{hdf(k)} \to 1. \tag{12}$$

Remark: If $\tilde{Y}_p$ is uniformly integrable, then $E(\tilde{Y}_p) \to 0$, and hence the result of Corollary 1.1 holds.

It can be seen that Corollary 1.1 holds given the assumptions, since both $\mathrm{hdf}(k)$ and $\mathrm{df}_C(k)$ diverge while $E(\tilde{Y}_p)$ and the remainder term remain bounded. The Corollary suggests that for large k and large p, the ratio of $\mathrm{df}_C(k)$ to $\mathrm{hdf}(k)$ will be close to 1. We next explore empirically the relative behavior of the two dfs for a fixed p with an increasing k.

2.1.4 Numerical justification of hdf

Figure 1 shows the comparison of hdf and edf via simulations. We fit BS on 1000 realizations of the response generated after fixing X. The edf is calculated based on definition (3) using the sample covariances, while hdf is given by Algorithm 1. We see that in the null case, using hdf to approximate edf becomes more accurate as k approaches p, providing a finite-sample justification of Corollary 1.1.

In addition to the null model, we consider a sparse model (Orth-Sparse-Ex1) with $p_0 = 6$ true predictors (those with non-zero coefficients), and a dense model (Orth-Dense) where all predictors have non-zero coefficients. We also consider two signal-to-noise (SNR) ratios with 'hsnr' and 'lsnr' representing high and low SNR respectively, and the SNR is defined as $\mathrm{Var}(x^T\beta)/\sigma^2$. The details of the setups for Orth-Sparse-Ex1 and Orth-Dense models can be found in Section 2.3.1. Similarly to the null case, we see that hdf approaches edf as k gets close to p, i.e. the statement of Corollary 1.1 holds in these scenarios as well. Furthermore, we see that hdf generally approximates edf well, where the difference is more pronounced when BS underfits, e.g. a sparse model with high SNR and $k < p_0 = 6$ or a dense model with high SNR with $k < p = 14$. Clearly, underfitting causes the problem, particularly when what is left out is important, such as in a high SNR case.

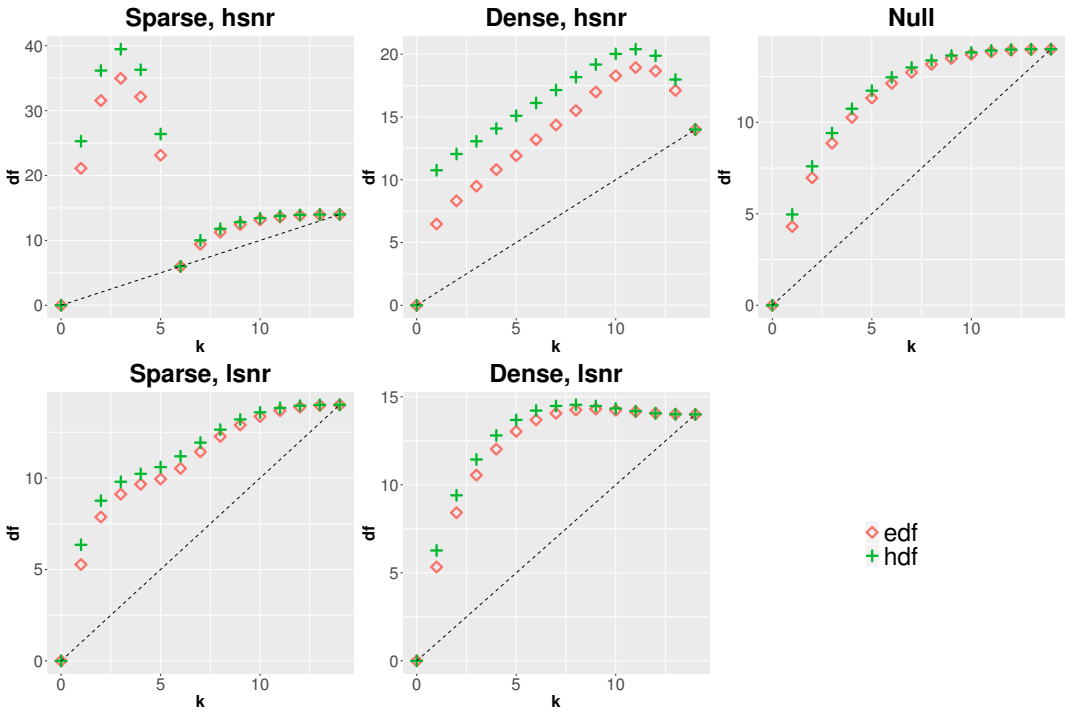

Figure 1: hdf(k) and df$_C(k)$ (edf of constrained BS). The black dashed line is the 45-degree line. Here X is orthogonal with $n = 200$ and $p = 14$. Three types of the true model and two SNR are considered. We assume knowledge of μ and σ.

2.1.5 C_p-hdf as a feasible implementation of C_p-edf

We have shown that hdf generally approximates edf well, and it agrees with edf as k approaches p. By replacing edf with hdf in (5), we have a feasible selection rule C_p-hdf. Figure 2 compares the averages of C_p-edf and C_p-hdf over 1000 replications. Similarly to the comparison of the degrees of freedom values, we see C_p-hdf agrees with C_p-edf on average as k approaches p. Even at the places where we see differences between the degrees of freedom values, e.g. a sparse true model with high SNR and $k < p_0 = 6$, the differences are compensated by the model fit and we see C_p-hdf is very close to C_p-edf. As we discussed in Section 1.2, for any general fitting procedure including BS, C_p-edf provides an unbiased estimator of the expected prediction error where the error measure Θ is the squared error (SE), i.e. $E(C_p\text{-edf}) = E(\text{Err}_{\text{SE}})$. Therefore, by using the sample average to represent the population mean, Figure 2 indicates that $E(C_p\text{-hdf})$ approximates $E(\text{Err}_{\text{SE}})$ well, and moreover C_p-hdf gives the same average selected size as C_p-edf in all cases, when they are applied as selection rules, supporting the use of hdf in model selection for BS.

2.2 A KL-based information criterion for BS

When the error measure Θ is the deviance (6), the prediction error Err_{KL} is the KL discrepancy. AICc-edf (7) is motivated by trying to construct an unbiased estimator of $E(\text{Err}_{\text{KL}})$. The expected KL-based optimism for BS is given as

$$E(\text{op}_{\text{KL}}) = E\left(n\frac{n\sigma^2 + \|\mu - X\hat{\beta}(k)\|_2^2}{\|y - X\hat{\beta}(k)\|_2^2}\right) + n. \tag{13}$$

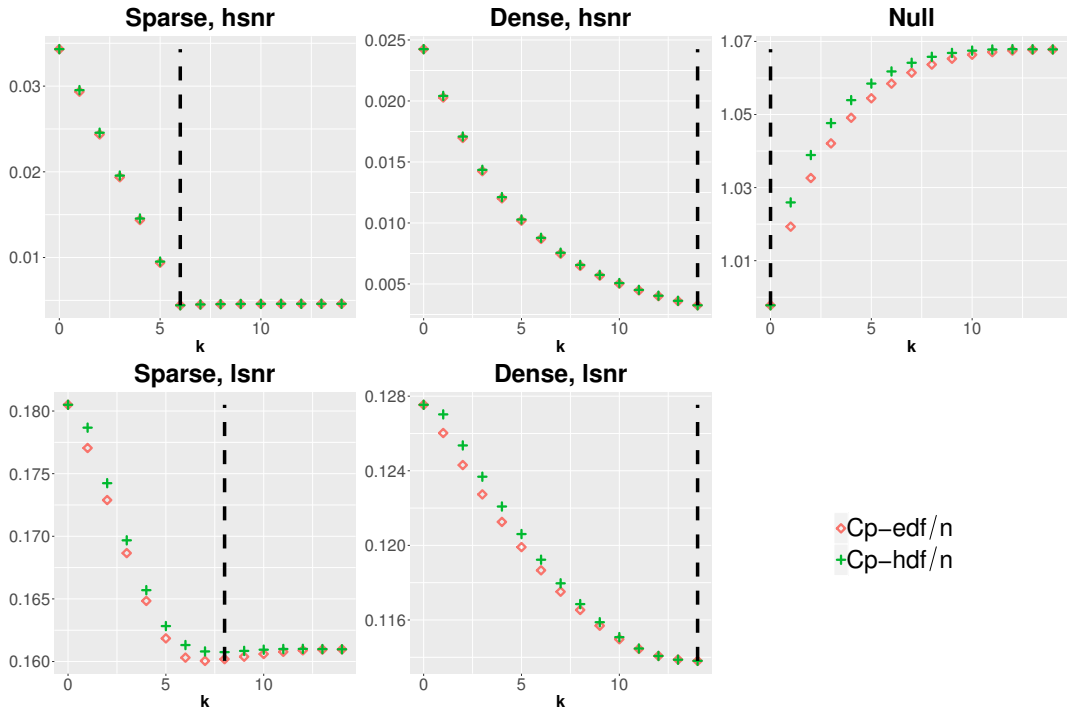

Figure 2: Averages of C_p-edf and C_p-hdf over 1000 replications. Both criteria lead to the same average of the selected subset size over the 1000 replications (rounded to the nearest integer), as denoted by the black dashed vertical lines. Other details are the same as in Figure 1.

Note that (13) holds for a general X. Augmenting $E(\text{op}_{\text{KL}})$ with the training error err_{KL} we have $\widehat{\text{Err}}_{\text{KL}}$ according to (4), where (ignoring the constant $n\log(2\pi)$ for convenience)

$$\text{err}_{\text{KL}} = n\log\left(\frac{\text{RSS}}{n}\right) - n, \tag{14}$$

since the pre-specified model f in (6) follows a Gaussian distribution as assumed in (1). The derivations of (13) and (14) are presented in the Supplemental Material.

Figure 3 shows the averages of AICc-edf, AICc-hdf (AICc-hdf is calculated by replacing edf with hdf in (7)) and $\widehat{\text{Err}}_{\text{KL}}$ over 1000 replications. By using the sample average to represent population mean, we first see that $E(\text{AICc-edf})$ generally tracks the expected KL, $E(\text{Err}_{\text{KL}})$ reasonably well. In fact, they agree with each other in the null case and a sparse true model with high SNR. Noticeable discrepancies can be observed in a sparse true model with high SNR and $k < p_0 = 6$. This is the place where the set of true predictors is not entirely included in the model. The derivations of the classic AIC and AICc (both with ndf plugged in according to our notation) are based on an assumption that the true predictors are included in the model. In the situation where this assumption is violated, AICc will no longer be unbiased, and a similar conjecture can be made here for AICc-edf in the context of BS. Second, similarly to the comparison of $E(C_p\text{-edf})$ and $E(C_p\text{-hdf})$ in Section 2.1.5, we see that $E(\text{AICc-hdf})$ approximates $E(\text{AICc-edf})$ well and they agree with each other as k approaches p. Last and most importantly, both AICc-edf and AICc-hdf yield the same average selected size as $\widehat{\text{Err}}_{\text{KL}}$ across all scenarios, supporting the use of AICc-hdf as a selection rule for BS.

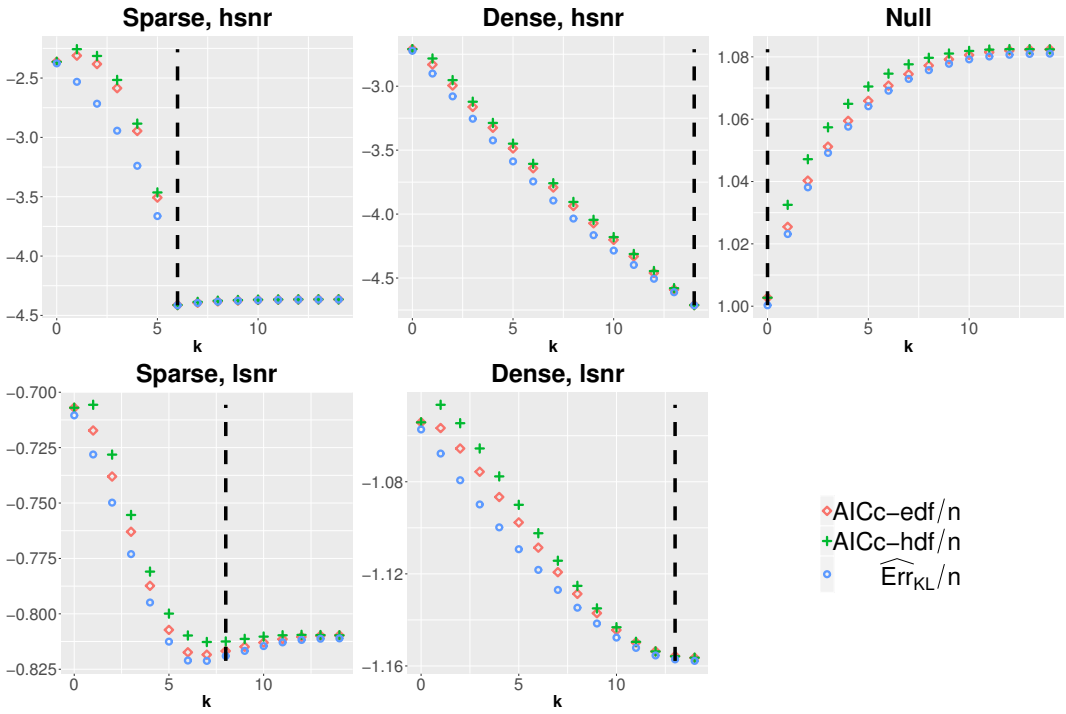

Figure 3: Averages of AICc-edf, AICc-hdf and $\widehat{\text{Err}}_{\text{KL}}$ over 1000 replications. All three criteria lead to the same average of the selected subset size over the 1000 replications (rounded to the nearest integer), as denoted by the black dashed vertical lines. Other details are the same as in Figure 1.

2.3 The performance of AICc-hdf as a selection rule for BS

We now study the performance of AICc-hdf as a selection rule for BS. The only assumption we make in the simulations is X being orthogonal. Both μ and σ are treated unknown, as would be the case in practice. We start by showing that AICc-hdf can perform well for BS compared to other selection rules. We then compare the performance of BS-AICc-hdf to that of regularization methods.

2.3.1 Simulation set-up

We consider a trigonometric configuration of X that is studied by Hurvich and Tsai (1991), where $X = (x^1, x^2)$ is an n by p matrix with components defined by

$$x^1_{tj} = \sin\left(\frac{2\pi j}{n}t\right),$$

and

$$x^2_{tj} = \cos\left(\frac{2\pi j}{n}t\right),$$

for $j = 1, \cdots, p/2$ and $t = 0, \cdots, n-1$. The columns of X are then standardized to have l_2 norm 1, to make them orthonormal. By fixing X, the responses are generated by (1), where $\mu = X\beta$. The error ϵ is also shifted to have mean 0, hence the intercept will be zero.

We consider the following configurations of the experiment:

- Sample size: $n \in \{200, 2000\}$.

- Number of predictors: $p \in \{14, 30, 60, 180\}$.
- Signal-to-noise ratio: SNR $\in \{0.2, 1.5, 7\}$ (low, medium and high). The average oracle R^2 (linear regression on the set of true predictors) corresponding to these three SNR values are roughly 20%, 50% and 90%, respectively.
- Coefficient vector β (Orth in the following denotes for orthogonal X):
  - Orth-Sparse-Ex1: $\beta = [1_6, 0_{p-6}]^T$
  - Orth-Sparse-Ex2: $\beta = [1, -1, 5, -5, 10, -10, 0_{p-6}]^T$
  - Orth-Dense (Taddy, 2017): $\beta_j = (-1)^j \exp(-\frac{j}{\kappa})$, $j = 1, \cdots, p$. $\kappa = 10$

In total, there are 72 different scenarios in the experiment. The full set of simulation results is presented in the Supplemental Material. In each scenario, 1000 replications of the response y are generated. A fitting procedure $\hat{\mu}$, is evaluated via the average RMSE, where

$$\text{RMSE}(\hat{\mu}) = \sqrt{\frac{1}{n}\|\hat{\mu} - X\beta\|_2^2}. \tag{15}$$

To make the scales easier to compare, we construct two relative metrics: % worse than the best possible BS, and relative efficiency, which are defined as follows:

- % **worse than best possible BS**

$$= 100 \times \left(\frac{\text{averge RMSE of a fitting procedure } \hat{\mu}}{\text{average RMSE of the best possible BS}} - 1 \right) \%, \tag{16}$$

  where the best possible BS here means that on a single fit, choosing the subset size with the minimum RMSE among all $p + 1$ candidates, as if an oracle tells us the true model.

- **Relative efficiency:** For a collection of fitting procedures, the relative efficiency for a particular procedure j, is defined as

$$\frac{\min_l \text{ average RMSE of fitting procedure } l}{\text{average RMSE of fitting procedure } j}. \tag{17}$$

  The relative efficiency is a measure between 0 and 1. Higher value indicates better performance. Besides the fitting procedures specified, we include the null and full OLS in the calculation of relative efficiency.

We also present the sparsistency (number of true positives) and number of extra predictors (number of false positives).

2.3.2 AICc-hdf and other selection rules for BS

By analogy to C_p and AICc, we can also define BIC-edf as

$$\text{BIC-edf} = n \log\left(\frac{\text{RSS}}{n}\right) + \log(n) \cdot \text{edf}, \tag{18}$$

and its feasible version BIC-hdf, where the original BIC (or BIC-ndf in our notation) was introduced in Schwarz (1978). We also consider a numerical estimation of edf that is based on the parametric bootstrap, and we denote it as bdf. The detailed implementation of bdf and the benefit of parametric bootstrap is discussed in Efron (2004). In our experiment, we use 100 bootstrapped samples. In addition to the information criteria, we further include 10-fold cross-validation (CV) for comparison. Note that the CV results are only available for $p \leq 30$ since it is fitted using the 'leaps' algorithm.

A selected set of results is shown in Tables 1 and 2. A brief summary is as follows:

- Using information criteria in the naive way (with ndf) can be dangerous, especially when p is large and SNR is high. For example, using ndf in AICc significantly overfits and can be almost 400 times worse in terms of RMSE than using hdf for $n = 200$, high SNR and $p = 180$ in Orth-Sparse-Ex1. Increasing the sample size n does not improve the naive implementation of information criteria, and the overfiting persists.

- AICc-hdf generally does not lose much efficiency and performs similarly in terms of RMSE, in comparison to the infeasible AICc-edf. Increasing the sample size n or SNR improves the performance of both AICc-edf and AICc-hdf.

- AICc-hdf performs very similarly to AICc-bdf. Since bdf is calculated based on 100 bootstrapped samples, it is roughly 100 times more intensive than hdf in computations.

- AICc-hdf is generally better than 10-fold CV, e.g. when n is large or SNR is high. Note that 10-fold CV is roughly 10 times heavier in terms of computation than AICc-hdf. It is also worth noticing that these findings are broadly consistent with the results reported by Taddy (2017) for the gamma lasso method.

- C_p-edf performs similarly to AICc-edf. In contrast, when we consider the feasible implementations (ndf/hdf/bdf), i.e. when σ is estimated by full OLS, C_p can suffer when p is close to n, such as when $n = 200$ and $p = 180$. Under a sparse true model BIC-hdf performs slightly better than AICc-hdf except when SNR is low and $n = 200$, where BIC is considerably worse. Under a dense true model BIC-hdf is always outperformed by AICc-hdf.

For the reasons presented above, we conclude that AICc-hdf is the best feasible selection rule for BS, among all that have been considered.

2.3.3 How does BS perform compared to regularization methods?

We have seen that AICc-hdf can be an effective selection rule for BS. In this section, we compare BS with some popular regularization methods, including lasso (Tibshirani, 1996), SparseNet (Mazumder et al., 2011), gamma lasso (Taddy, 2017) and relaxed lasso (Meinshausen, 2007). We use `R` packages **glmnet** (Friedman et al., 2010), **sparsenet** (Mazumder et al., 2011), **gamlr** (Taddy, 2017) and **relaxo** (Meinshausen, 2012), to fit them, respectively, which are all available on *CRAN*.

As to the selection rule, we use AICc-hdf for BS, AICc for lasso, and 10-fold CV for the rest. In addition to these selectors, we have also considered 10-fold CV for lasso. We find (in the Supplement) that 10-fold CV performs similarly to AICc for lasso. In fact, the use of AICc for lasso has been explored in Flynn et al. (2013), where the authors proved that AICc is asymptotically efficient while performing similarly to CV. We further notice (results given in the Supplement) that SparseNet generally performs better than the relaxed lasso and gamma lasso, and hence only the results for SparseNet will be presented here.

A selected set of results is presented in Table 3. A brief summary is as follows:

- For a relatively small sample size $n = 200$ and a sparse true model, BS performs the best when SNR is high, lasso is best in low SNR, and SparseNet has performance in between of the other two methods. lasso has the property of 'over-select and shrink,' in order to retain less bias on the large non-zero estimates. In a high SNR, this property can result in disastrous performance, especially when p is close to n. For example, in Orth-Sparse-Ex1, high SNR and $p = 180$, the relative efficiency of lasso is only 0.44 and it significantly overfits. However, this property can be beneficial when SNR is low, as a method like BS has higher chance to miss the true predictors (less sparsistency).

- With $n = 200$ and a dense true model, the methods perform similarly when SNR is high, while lasso is better in low SNR.

Table 1: The performance of AICc-hdf. The true model setup is Orth-Sparse-Ex1. The columns involving 'edf' refer to infeasible selection rules since edf is estimated as if the true model is known, while other columns correspond to feasible rules.

			C_p		AICc		BIC		CV
			edf	ndf/hdf/bdf	edf	ndf/hdf/bdf	edf	ndf/hdf/bdf	
			% worse than the best possible BS						
n=200	hsnr	p=30	4	84/5/7	2	83/2/5	0	28/0/0	24
		p=180	1	338/30/32	0	392/1/2	0	206/0/0	-
	lsnr	p=30	20	25/36/33	21	24/37/35	68	23/68/67	28
		p=180	15	108/35/34	18	132/22/22	25	50/25/25	-
n=2000	hsnr	p=30	3	85/3/6	3	85/3/6	0	9/0/0	23
		p=180	0	334/1/3	1	337/1/3	0	60/0/0	-
	lsnr	p=30	3	85/6/7	3	85/5/6	0	9/0/0	23
		p=180	0	334/5/4	1	337/4/4	0	60/1/1	-
			Relative efficiency						
n=200	hsnr	p=30	0.96	0.54/0.95/0.93	0.98	0.55/0.98/0.96	1	0.78/1/1	0.81
		p=180	0.99	0.23/0.77/0.76	1	0.2/0.99/0.98	1	0.33/1/1	-
	lsnr	p=30	1	0.97/0.89/0.9	1	0.97/0.88/0.89	0.72	0.98/0.72/0.72	0.94
		p=180	1	0.55/0.86/0.86	0.97	0.5/0.95/0.95	0.93	0.77/0.93/0.93	-
n=2000	hsnr	p=30	0.97	0.54/0.97/0.94	0.97	0.54/0.97/0.94	1	0.92/1/1	0.81
		p=180	1	0.23/0.99/0.97	0.99	0.23/0.99/0.97	1	0.62/1/1	-
	lsnr	p=30	0.97	0.54/0.95/0.94	0.97	0.54/0.95/0.94	1	0.92/1/1	0.81
		p=180	1	0.23/0.96/0.96	0.99	0.23/0.96/0.96	1	0.62/0.99/0.99	-
			Sparsistency (number of extra variables)						
n=200	hsnr	p=30	6(0.1)	6(3.9)/6(0.2)/6(0.2)	6(0)	6(3.8)/6(0.1)/6(0.1)	6(0)	6(0.6)/6(0)/6(0)	6(0.7)
		p=180	6(0)	6(32.2)/6(6.4)/6(6.3)	6(0)	6(48.9)/6(0)/6(0)	6(0)	6(9.5)/6(0)/6(0)	-
	lsnr	p=30	4.5(1.9)	5.3(3.9)/4.2(4.9)/4.2(4)	4.2(1.2)	5.2(3.8)/3.3(2.2)/3.4(1.8)	0.1(0)	3.7(0.6)/0.1(0)/0.2(0)	4(1.9)
		p=180	1.9(0.5)	5.3(32.2)/1.8(10.9)/1.9(9.8)	1.1(0.1)	5.6(49)/0.5(0)/0.6(0)	0(0)	4.2(8.4)/0(0)/0(0)	-
n=2000	hsnr	p=30	6(0.1)	6(3.8)/6(0.1)/6(0.2)	6(0.1)	6(3.8)/6(0.1)/6(0.2)	6(0)	6(0.1)/6(0)/6(0)	6(0.6)
		p=180	6(0)	6(27.5)/6(0)/6(0)	6(0)	6(28.2)/6(0)/6(0)	6(0)	6(1.1)/6(0)/6(0)	-
	lsnr	p=30	6(0.1)	6(3.8)/6(0.2)/6(0.2)	6(0.1)	6(3.8)/6(0.2)/6(0.2)	6(0)	6(0.1)/6(0)/6(0)	6(0.6)
		p=180	6(0)	6(27.5)/6(0.1)/6(0)	6(0)	6(28.2)/6(0.1)/6(0)	6(0)	6(1.1)/6(0)/6(0)	-

- With a large sample size $n = 2000$ relative to the values of p, BS becomes the best in almost all scenarios. The only exception is when the true model is dense and SNR is low, where BS is very close to the best. In fact, all three methods benefit from increasing n, since we can see larger sparsistency and fewer extra variables. Given that, it seems that BS profits the most according to the boost of its relative performance in low SNR.

Given the spirit of the summary above, it's important to point out the relevant work of Hastie et al. (2017), where the authors provide a comprehensive set of simulation comparisons on BS, lasso and relaxed lasso. The authors concluded that BS performs the best in high SNR, lasso is the best in low SNR while relaxed lasso is in between. This coincides with the results here when sample size is relatively small ($n = 200$), given the similarity in the performance of relaxed lasso and SparseNet. However, we find BS to be the best for large sample size n even when the SNR is low (note that Hastie et al. (2017) did not examine any sample sizes greater than $n = 500$). Moreover, it should be noted that Hastie et al. (2017) focus on the best possible performance of each method by applying a separate validation set drawn from the true model, rather than on feasible selection, as is considered in this study.

Table 2: The performance of AICc-hdf. The true model setup is Orth-Dense. Details of the columns can be referred to the caption in Table 1

			C_p		AICc		BIC		CV
			edf	ndf/hdf/bdf	edf	ndf/hdf/bdf	edf	ndf/hdf/bdf	
			% worse than the best possible BS						
n=200	hsnr	p=30	1	11/1/2	1	13/1/2	1	28/3/5	7
		p=180	7	45/21/20	9	52/18/19	18	26/39/42	-
	lsnr	p=30	15	10/16/16	20	10/21/20	27	16/27/27	16
		p=180	8	86/22/22	7	102/7/7	7	39/7/7	-
n=2000	hsnr	p=30	0	1/0/0	0	1/0/0	0	18/0/1	1
		p=180	6	34/8/8	6	34/8/8	19	7/36/37	-
	lsnr	p=30	2	11/3/3	2	11/3/3	44	41/36/45	10
		p=180	8	48/10/10	8	48/10/10	24	8/45/47	-
			Relative efficiency						
n=200	hsnr	p=30	1	0.91/1/1	1	0.9/1/0.99	1	0.79/0.98/0.96	0.95
		p=180	1	0.74/0.89/0.89	0.99	0.71/0.91/0.9	0.91	0.85/0.77/0.76	-
	lsnr	p=30	0.95	1/0.95/0.95	0.91	1/0.91/0.91	0.86	0.94/0.86/0.86	0.94
		p=180	1	0.58/0.88/0.88	1	0.53/1/1	1	0.77/1/1	-
n=2000	hsnr	p=30	1	0.99/1/1	1	0.99/1/1	1	0.85/1/0.99	0.99
		p=180	1	0.79/0.98/0.98	1	0.79/0.98/0.98	0.89	1/0.78/0.78	-
	lsnr	p=30	1	0.92/0.99/0.99	1	0.92/0.99/0.99	0.71	0.73/0.75/0.7	0.93
		p=180	1	0.73/0.98/0.98	1	0.73/0.98/0.98	0.87	1/0.74/0.73	-
			Sparsistency (number of extra variables)						
n=200	hsnr	p=30	30	24.7/29.5/29	30	24.2/29.4/28.8	30	20.9/28.8/27.5	26.6
		p=180	20.5	53.3/37.4/35.5	18.3	62.3/16.3/16.3	16.1	35/13.7/13.5	-
	lsnr	p=30	12.8	10.5/14.6/13	7.6	10.3/8.5/7.6	0	4/0/0	7.5
		p=180	0.8	39/14.5/13.7	0.3	55.2/0.2/0.3	0	11.8/0/0	-
n=2000	hsnr	p=30	30	29.8/30/29.9	30	29.8/30/29.9	30	28.6/30/29.9	29.8
		p=180	32.1	58.9/32.4/32.3	31.8	58.9/31.6/31.6	27	31.3/25/24.9	-
	lsnr	p=30	28.8	19.9/28.2/26.9	28.8	19.9/28.1/26.8	13.5	12.5/16.7/14.1	22.3
		p=180	13.9	43.8/14/14	13.6	44.1/13.3/13.3	9.1	13.4/7/6.8	-

2.4 A discussion on the use of information criteria in LBS

Since for orthogonal X the edf of LBS has an analytical expression and LBS can recover the solution path of BS, one may ask why not just use LBS with a selection rule such as C_p-edf, which is well-defined for any general fitting procedure.

We consider a fixed sequence of λ and compute the LBS solutions for 1000 realizations. The decreasing sequence of λ starts at the smallest value $\lambda_{\max}$ for which the estimated coefficient vector $\hat{\beta}$ equals zero for all of the 1000 realizations. We then construct a sequence of 200 values of λ equally spaced in log scale from $\lambda_{\max}$ to $\lambda_{\min}$, where $\lambda_{\min} = \alpha\lambda_{\max}$ and $\alpha = 0.001$. This procedure of generating the sequence of λ has been discussed by Friedman et al. (2010) in the context of lasso.

Table 4 shows that LBS is outperformed by BS in almost all scenarios based on 1000 simulation replications. We use C_p-edf as the selection rule for both methods, where edf of BS ($\mathrm{df}_C(k)$) is estimated via simulations and edf of LBS ($\mathrm{df}_L(\lambda)$) is calculated using formulas (9) and (10). We see that 1) LBS deteriorates as p gets larger when SNR is low or sample size n is large; and 2) increasing the sample

Table 3: The performance of BS compared to regularization methods. The selection rules are AICc-hdf for BS, AICc for lasso and 10-fold CV for SparseNet, respectively.

			Orth-Sparse-Ex1			Orth-Sparse-Ex2			Orth-Dense		
			BS	lasso	SparseNet	BS	lasso	SparseNet	BS	lasso	SparseNet
			% worse than the best possible BS								
n=200	hsnr	p=30	2	70	14	25	32	19	1	1	4
		p=180	1	128	20	11	51	14	18	23	7
	lsnr	p=30	37	7	15	34	21	25	21	-6	-2
		p=180	22	1	8	39	24	25	7	-2	2
n=2000	hsnr	p=30	3	70	14	5	71	15	0	1	1
		p=180	1	130	14	4	129	14	8	17	3
	lsnr	p=30	5	70	15	13	46	15	3	1	5
		p=180	4	129	14	8	88	15	10	9	5
			Relative efficiency								
n=200	hsnr	p=30	1	0.6	0.9	0.95	0.9	1	1	1	0.97
		p=180	1	0.44	0.84	1	0.74	0.97	0.91	0.87	1
	lsnr	p=30	0.78	1	0.93	0.9	1	0.97	0.78	1	0.95
		p=180	0.83	1	0.94	0.9	1	0.99	0.91	1	0.96
n=2000	hsnr	p=30	1	0.61	0.91	1	0.62	0.91	1	0.99	0.99
		p=180	1	0.44	0.89	1	0.45	0.91	0.96	0.88	1
	lsnr	p=30	1	0.62	0.92	1	0.77	0.98	0.98	1	0.96
		p=180	1	0.46	0.91	1	0.58	0.95	0.95	0.96	1
			Sparsistency (number of extra variables)								
n=200	hsnr	p=30	6(0.1)	6(7.7)	6(1.1)	4.5(0.2)	5.7(7.2)	5.1(2)	29.4	28.9	27.8
		p=180	6(0)	6(14.3)	6(3.7)	4.1(0)	5.4(12.9)	4.7(5.3)	16.3	41.1	32.7
	lsnr	p=30	3.3(2.2)	5.3(6.7)	4.9(5.1)	2.5(0.9)	3.9(5.1)	3.1(2.7)	8.5	14.1	13.3
		p=180	0.5(0)	3.9(9.6)	3.6(11.3)	1.1(0.1)	3(9.6)	2.6(7.7)	0.2	10.5	13.6
n=2000	hsnr	p=30	6(0.1)	6(8.5)	6(1.1)	6(0.2)	6(8.5)	6(0.9)	30	30	29.9
		p=180	6(0)	6(21.8)	6(2.2)	6(0)	6(21.5)	6(1.4)	31.6	89.1	46.3
	lsnr	p=30	6(0.2)	6(8.5)	6(0.8)	4.2(0.4)	5.1(7.3)	4.3(1.2)	28.1	26.8	24.8
		p=180	6(0.1)	6(21.2)	6(1.3)	4(0.1)	4.7(18.1)	4.1(1.5)	13.3	55.1	29.3

size n does not help LBS. Figure 4 further compares the number of predictors given by BS and LBS for each of the 1000 replications, where we consider the Orth-Sparse-Ex1 model with $n = 200$ and high SNR. Under this design, LBS never selects fewer predictors than BS, and it chooses more predictors in 31.1% and 37.8% of all replications for $p = 30$ and $p = 180$, respectively. A possible explanation for this might be that $\mathrm{df}_L(\lambda)$ characterizes the model complexity at λ on average, but does not correctly describe the model complexity on a given realization. Given a single realization, there are an infinite number of λ values that correspond to each distinct solution, and they lead to different values of $\mathrm{df}_L(\lambda)$ and further result in different model complexities and different C_p values. This variability in the C_p values for the same solution potentially causes the selected subsets of LBS to have more variabilities than those selected subsets of BS.

Table 4: The performance of BS and LBS. The selection rule for both methods is C_p-edf. We assume knowledge of μ and σ.

			Orth-Sparse-Ex1		Orth-Sparse-Ex2		Orth-Dense	
			BS	LBS	BS	LBS	BS	LBS
			% worse than the best possible BS					
n=200	hsnr	p=30	4	28	21	25	1	1
		p=180	1	43	12	25	7	10
	lsnr	p=30	20	26	21	30	15	16
		p=180	15	20	15	32	8	12
n=2000	hsnr	p=30	3	29	3	28	0	0
		p=180	0	42	0	41	6	9
	lsnr	p=30	3	28	7	25	2	2
		p=180	0	41	1	37	8	12
			Relative efficiency					
n=200	hsnr	p=30	1	0.81	1	0.96	1	1
		p=180	1	0.71	1	0.89	1	0.97
	lsnr	p=30	1	0.96	1	0.93	0.95	0.94
		p=180	1	0.96	1	0.87	1	0.95
n=2000	hsnr	p=30	1	0.8	1	0.81	1	1
		p=180	1	0.71	1	0.71	1	0.98
	lsnr	p=30	1	0.81	1	0.86	1	1
		p=180	1	0.71	1	0.74	1	0.97
			Sparsistency (number of extra variables)					
n=200	hsnr	p=30	6(0.1)	6(0.7)	4.8(0.4)	5(0.9)	30	29.9
		p=180	6(0)	6(0.9)	4.2(0)	4.6(0.9)	20.5	21.9
	lsnr	p=30	4.5(1.9)	4.4(2.3)	2.7(0.6)	2.9(1.1)	12.8	7.6
		p=180	1.9(0.5)	2.3(1.4)	1.9(0.3)	2.2(1.2)	0.8	2.7
n=2000	hsnr	p=30	6(0.1)	6(0.7)	6(0.1)	6(0.7)	30	30
		p=180	6(0)	6(0.8)	6(0)	6(0.8)	32.1	34.1
	lsnr	p=30	6(0.1)	6(0.7)	4.1(0.2)	4.3(0.7)	28.8	29.4
		p=180	6(0)	6(0.8)	4(0)	4.1(0.8)	13.9	15.9

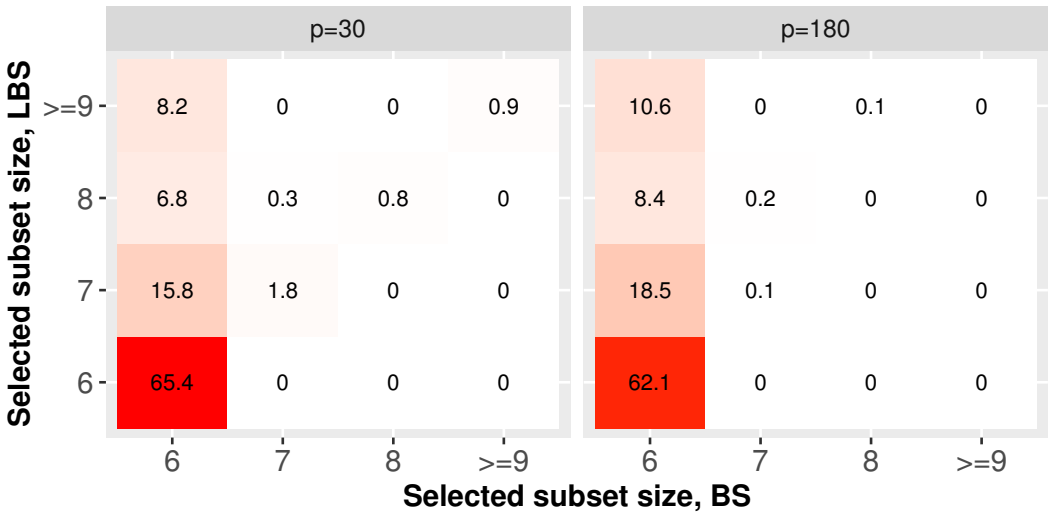

Figure 4: Frequency distributions (in %) of the selected subset size given by BS and LBS, based on 1000 replications. The selection rule is C_p-edf. The true model is Orth-Sparse-Ex1 with $n = 200$, $p_0 = 6$ and high SNR.

3 Best orthogonalized subset selection (BOSS)

With a general X, BS is not computationally feasible for large problems. In this section, we propose a LS-based subset selection method BOSS that has computational cost of the same order as a multiple regression on all of the predictors.

3.1 The method and its computational cost

The detailed implementation process of BOSS is described in Algorithm 2. The main steps can be summarized as follows: 1) order and orthogonalize the predictors, 2) perform BS on the set of orthogonal predictors, 3) transform the coefficients back to the original space, and 4) use a selection rule such as AICc-hdf to choose the single subset.

As can be seen in what follows, the computation of BOSS has an overall cost of $O(np^2)$, the same cost as OLS on all p predictors and lasso. For step k, we have a set of ordered predictors $X_{S_{k-1}}$ and its orthogonal basis $Q_{S_{k-1}}$. From the remaining $p-k+1$ predictors, we choose the one that has the largest correlation with y conditioning on $Q_{S_{k-1}}$, that is the correlation between y and the residual from regressing a candidate predictor on $Q_{S_{k-1}}$. The regression part costs $O(n)$ since we maintain the regression result, e.g. estimated coefficients and residual, in the previous steps, and only need to perform a simple linear regression upon the predictor joined in step $k-1$, i.e. the last column in $Q_{S_{k-1}}$. Repeating the above step for all $p-k+1$ predictors costs $O(n(p-k+1))$. We then update the QR decomposition, by adding the chosen predictor as a new column, which costs $O(n(p-k))$ via the modified Gram-Schmidt algorithm as discussed in Hammarling and Lucas (2008). Therefore, we end up with an ordered set of predictors X_{S_p} and its corresponding QR decomposition Q_{S_p} and R_{S_p}. We regress y upon Q_{S_p} which costs $O(np)$, and denote the coefficient vector as z. BOSS then performs BS on Q_{S_p}, which is a ranking of predictors based on their magnitudes of corresponding element in z, and the cost is $O(p\log(p))$. Once we have the solution path of BOSS, we then apply AICc-hdf to choose

Algorithm 2 Best Orthogonalized Subset Selection (BOSS)

1. Standardize y and the columns of X to have mean 0, and denote the means as $\bar{X}$ and $\bar{y}$.

   Order and orthogonalize the predictors:

2. From the p predictors, select the one that has the largest marginal correlation with the response y, and denote it as X_{S_1}. Standardize X_{S_1} to have unit l_2 norm and denote it as Q_{S_1}. Calculate R_{S_1} such that $X_{S_1} = Q_{S_1} R_{S_1}$. Let $S = \{1, \cdots, p\}$. Initialize vectors $\text{resid}_j = X_j$ where $j = 1, \cdots, p$.

3. For $k = 2, \cdots, p$:

   a. For each of the $p-k+1$ predictors X_j in $X_{S \setminus S_{k-1}}$, calculate its partial correlations with the response y conditioning on $Q_{S_{k-1}}$.

      a1. Regress X_j on $Q_{S_{k-1} \setminus S_{k-2}}$ ($S_{k-2} = \emptyset$ if $k = 2$), and denote the estimated coefficient as r. Update $\text{resid}_j = \text{resid}_j - r Q_{S_{k-1} \setminus S_{k-2}}$.

      a2. Calculate the correlation between y and resid_j.

   b. Select the predictor that has the largest partial correlation in magnitude, augment S_{k-1} with this predictor and call it S_k.

   c. Update $Q_{S_{k-1}}$ and $R_{S_{k-1}}$ given the newly added column $X_{S_k \setminus S_{k-1}}$, and call them Q_{S_k} and R_{S_k}. The update is based on the modified Gram-Schmidt algorithm as discussed in Hammarling and Lucas (2008).

   BS on the orthogonalized predictors Q_{S_p}:

4. Calculate $\tilde{\gamma}_j(k_Q) = z_j \mathbb{1}(|z_j| \geq |z_{(k_Q)}|)$, i.e. the j-th component of coefficient vector for subset size k_Q, where $z = Q_{S_p}^T y$ and $z_{(k_Q)}$ is the k-th largest entry in absolute values. Let $\tilde{\gamma} = [\tilde{\gamma}(0)\tilde{\gamma}(1) \cdots \tilde{\gamma}(p)]$.

   Transform back to the original space:

5. Project $\tilde{\gamma}$, a $p \times (p+1)$ matrix, to the original space of X_{S_p}, i.e. back solving $R\tilde{B} = \tilde{\gamma}$, and re-order the rows of $\tilde{B}$ to their correspondences in X, i.e. $\hat{B} = O\tilde{B}$ where O represents the ordering matrix s.t. $X_{S_p} = XO$. The intercept vector is $\hat{B}_0 = \bar{y}\mathbb{1} - \hat{B}^T \bar{X}$.

   Select the subset:

6. Select the subset using AICc-hdf, where hdf is calculated via Algorithm 1, by inputting (Q_{S_p}, y). The inclusion of an intercept term implies that $\text{hdf}(k_Q)$ is increased by 1.

the single subset size (denoted as k_Q), where hdf is calculated via Algorithm 1 by inputting Q_{S_p}. The entire BOSS-AICc-hdf procedure costs $O(np^2)$.

The ordering of predictors is essential in terms of both getting a sparse solution and a better predictive performance. Consider a sparse true model with only two uncorrelated predictors $X = [X_1, X_2]$, $\beta = [0, 1]^T$ and a high SNR. Based on the evidence we see from the previous section, the best model in such a scenario is LS regression on X_2. Without the ordering step, the orthogonal basis is $Q = [Q_1, Q_2]$ s.t. $X = QR$, i.e. the predictors are orthogonalized in their physical orders. The one-predictor model ($k_Q = 1$) of BS can either be Q_1 or Q_2, which when transformed back to the space of X do not correspond to LS regression upon X_2. The former corresponds to LS estimates upon X_1, while the latter is a linear combination of LS estimates upon X and LS estimates upon X_1; the former leads to a completely wrong model while the latter results in non-zero coefficients on both predictors. In contrast, if X_2 is the first variable orthogonalized, the best subset will be based on that variable alone, the correct choice. Therefore, the ordering step is crucial to both sparsity as well as predictive performance. Note that we show the coefficients of BOSS can be expressed as a linear combination of LS coefficients on subsets of X in Theorem 2 and the proof can be found in the Supplemental Material.

BS on the set of orthogonalized predictors gives the chance for BOSS to 'look back' at the predictors that are already stepped in. One may notice that BOSS is similar to forward stepwise regression (FS), which was first introduced in Efroymson (1960). FS orders and orthogonalizes the predictors in the same way as BOSS. It then takes the nested subsets $Q_{S_1}, Q_{S_2}, \cdots, Q_{S_p}$ as the candidate subsets and performs LS regression upon them. Therefore, once a predictor is stepped in, it remains in the subsets of every following step of FS. As we will show in the next section, under certain circumstances, FS can easily overfit, since noise predictors (those with $\beta_j = 0$) step in during early steps. However, BOSS revisits the predictors that have already stepped in and allows them to be dropped, resulting in a better predictive performance than FS.

Theorem 2 (Coefficients of BOSS are a linear combination of LS coefficients on subsets of X)**.** *Suppose X has full column rank and the columns are already ordered. $X = QR$ where Q is an $n \times p$ matrix with orthonormal columns and R is a $p \times p$ upper-triangular matrix. Let $S_k = \{j_1, j_2, \cdots, j_{k_Q}\}$ denote the support (position of predictors) of the best k_Q-predictor model given by BS upon (Q, y), and use $\hat{\gamma}(k_Q)$ (p by 1) to denote the BS coefficients. The corresponding coefficients in the X space, i.e. $\hat{\beta}(k_Q)$ s.t. $R\hat{\beta}(k_Q) = \hat{\gamma}(k_Q)$, can be expressed as*

$$\hat{\beta}(k_Q) = \sum_{j \in S_k} \left(\hat{\alpha}^{(j)} - \hat{\alpha}^{(j-1)} \right),$$

where the first j entries in $\hat{\alpha}^{(j)}$ (p by 1) are LS coefficients of regressing y upon $[X_1, X_2, \cdots, X_j]$ (the first j columns in X), and the remaining $p - j$ entries are zero.

3.2 Numerical justification of using hdf for BOSS

The hdf is designed for BS on a set of orthogonal predictors. However, before performing BS on the orthogonal basis, BOSS first orders the predictors. This raises the question as to whether hdf is reasonable to use in the selection rules for BOSS.

Figure 5 compares averages of C_p-edf and C_p-hdf over 1000 replications for BOSS under various true models. The details of setups for the sparse and dense models can be found in Section 3.3.1, where they correspond to the Sparse-Ex3 and Dense designs, respectively. The correlation between predictors is $\rho = 0.5$. We see that by using the sample average to represent the population mean, $E(C_p\text{-hdf})$ approximates $E(C_p\text{-edf})$ well in most cases where the latter is also the expected prediction error $E(\text{Err}_{\text{SE}})$ by definition, and they converge as k_Q approaches p. Moreover, both C_p-edf and C_p-hdf lead to the same average selected subset size, supporting the use of hdf in model selection for BOSS.

Note that for any general fitting procedure including BOSS, C_p provides an unbiased estimator of the testing error, and for that reason it is used in our discussions about the best-case performance when

σ is assumed to be known. Unfortunately, as discussed in Section 2.3.2, using C_p as the selection rule for BS can perform poorly in practice because of the need to estimate σ, particularly when p is close to n. A similar property of using C_p as the selection rule for BOSS will be shown in Section 3.3.3. Therefore we prefer AICc in feasible versions of selection since it can perform considerably better.

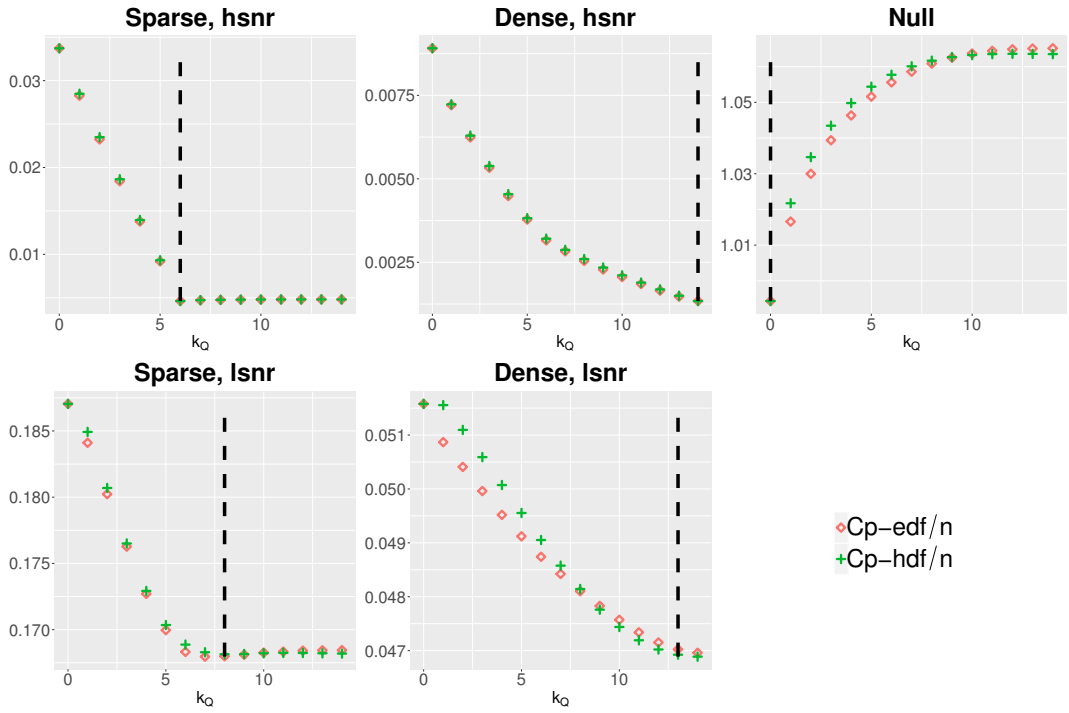

Figure 5: Averages of C_p-edf and C_p-hdf for BOSS over 1000 replications. Here X is general with $n = 200$, $p = 14$. Both criteria result in the same average of the selected subset size over the 1000 replications (rounded to the nearest integer) as denoted by the dashed vertical lines. We assume knowledge of μ and σ.

3.3 The performance of BOSS

We now study the performance of BOSS via simulations. We first show that BOSS can provide a better solution path than FS, and we further compare BOSS with regularization methods.

3.3.1 Simulation setups

We consider a similar setup as in Section 2.3.1, but with a general X, where $x_i \sim \mathcal{N}(0, \Sigma)$, $i = 1, \cdots, n$ are independent realizations from a p-dimensional multivariate normal distribution with mean zero and covariance matrix $\Sigma = (\sigma_{ij})$.

The correlation structure and true coefficient vector β include the following scenarios:

- Sparse-Ex1: **All of the predictors (both signal and noise) are correlated.** We take $\sigma_{i,j} = \rho^{|i-j|}$ for $i, j \in \{1, \cdots, p\} \times \{1, \cdots, p\}$. As to β, we have $\beta_j = 1$ for p_0 equispaced values and 0 everywhere else.

- Sparse-Ex2: **Signal predictors are pairwise correlated with opposite effects.** We take $\sigma_{i,j} = \sigma_{j,i} = \rho$ for $1 \leq i < j \leq p_0$. Other off-diagonal elements in Σ are zero. For the true

coefficient vector, we have $\beta_{2j-1} = 1$ and $\beta_{2j} = -1$ for $1 \leq j \leq p_0/2$, and all other $\beta_j = 0$ for $j = p_0 + 1, \cdots, p$.

- Sparse-Ex3: **Signal predictors are pairwise correlated with noise predictors.** We take $\sigma_{i,j} = \sigma_{j,i} = \rho$ for $1 \leq i \leq p_0$ and $j = p_0 + i$. Other off-diagonal elements in Σ are zero. $\beta = [1_{p_0}, 0_{p-p_0}]^T$.
- Sparse-Ex4: **Same correlation structure as Sparse-Ex2, but with varying strengths of coefficients.** We have $\beta_j = -\beta_{j+1}$ where $j = 2k + 1$ and $k = 0, 1, \cdots, p_0/2 - 1$. Suppose that $\beta' = [1, 5, 10]$, then $\beta_j = \beta'_k$ where $k = j(\text{mod}3)$.
- Dense: **Same correlation structure as Ex1, but with diminishing strengths of coefficients**. The true coefficient vector has: $\beta_j = (-1)^j \exp(-\frac{j}{\kappa})$, $j = 1, \cdots, p$, and here $\kappa = 10$.

The setup of Sparse-Ex1 is very common in the literature, such as in Bertsimas et al. (2016) and Hastie et al. (2017). All of the predictors are correlated (when $\rho \neq 0$) where the strength of correlation depends on the physical positions of variables. Sparse-Ex2 is designed such that the pair of correlated predictors, e.g. (X_1, X_2), leads to a good fit (high R^2), while either single one of them contribute little to the fitted R^2. Sparse-Ex4 is similar to Sparse-Ex2, but has varying strengths of coefficients for the true predictors. In Sparse-Ex3, signal predictors are only correlated with the noise ones. Finally, the dense setup is built on the dense example in Section 2.3.1, by having correlated predictors.

For the sparse examples, we take $p_0 = 6$. We consider three values of the correlation parameter, $\rho \in [0, 0.5, 0.9]$. Other configuration options, including n, p, and SNR, are the same as in Section 2.3.1. This implies a total of 360 different combinations of configuration options. For each configuration, 1000 replications are estimated and we present the same evaluation measures as introduced in Section 2.3.1. The full set of results can be found in the Supplemental Material.

3.3.2 The solution paths of BOSS and FS

Unlike FS, whose candidate subsets are nested, BOSS performs an extra step of BS upon Q_{S_p}, which raises the question of whether the extra step brings any benefit. We set aside the selection rule for now, and focus on the solution paths of the two methods.

Figure 6 shows two examples of the average RMSE along the solution paths of BS, FS and BOSS. When the true model is Sparse-Ex3, all three methods provide almost the same solution path. However, for Sparse-Ex4, we see a clear advantage of BOSS over FS in early steps up until about the fifteenth step. Recall that in Sparse-Ex4, there are $p_0 = 6$ predictors with $\beta_j \neq 0$ that are pairwise correlated with opposite effects, where each pair say (X_1, X_2) together leads to a high R^2 but each single one of them (X_1 or X_2) contributes little. When the correlation between X_1 and X_2 is high, the effect of X_1 almost completely cancels out the effect of X_2 on y. Therefore all of predictors (both true and noise predictors) have approximately zero marginal correlation with y, and they have equal chance of stepping in. Since the subsets along the solution path of FS are nested, if a noise predictor steps in during early steps, it remains in the subsets of every following step, and hence the subset containing both X_1 and X_2 may appear in a late stage. In contrast, BOSS takes ordered predictors provided by FS, and re-orders them by performing BS upon their orthogonal basis, which gives a greater chance for (X_1, X_2) to appear early in the solution path of BOSS and potentially results in a better predictive performance than FS. Furthermore, in this example, we notice that BOSS provides a better solution path than BS until step 5 (except the fourth step), and the two methods give similar performances in further steps.

3.3.3 The performance of BOSS compared to other methods

We now consider feasible implementations of the estimation methods. We looked at results using AICc-hdf, C_p-hdf and 10-fold CV for BOSS, and AICc-hdf was the best (see Supplemental Material),

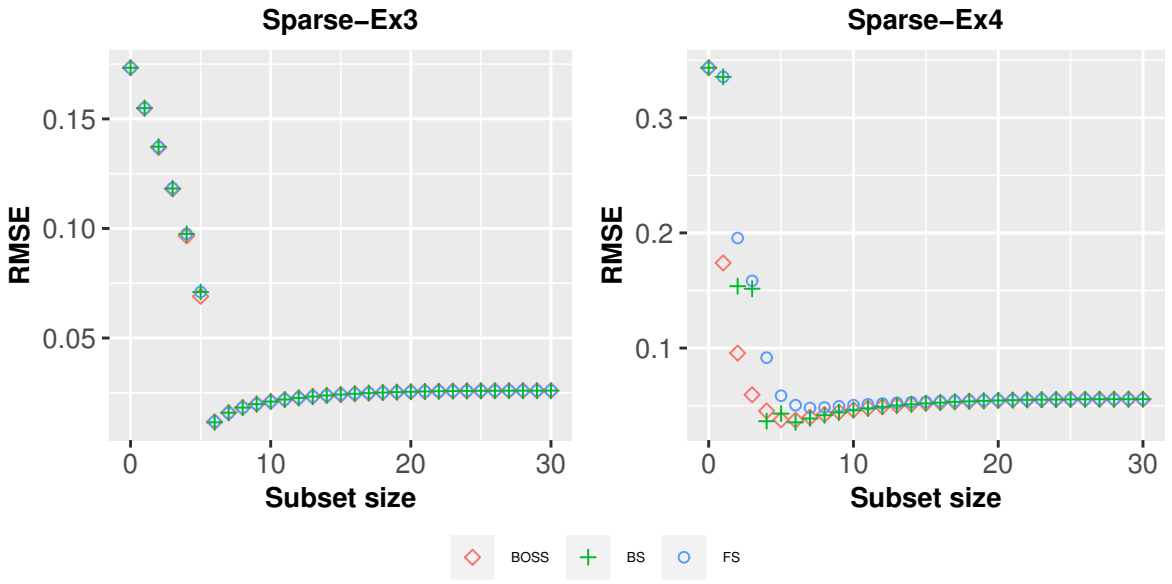

Figure 6: RMSE at each subset size, average over 1000 replications. Note that for BOSS, the subset size k_Q denotes the number of non-zero coefficients in $\hat{\gamma}(k_Q)$. In both scenarios, we have $n = 200$, $p = 30$, $\rho = 0.9$ and high SNR.

so that is what we will use here. For BS and FS, we will use 10-fold CV. Similar to our discussion in Section 2.3.3, we find that (see Supplemental Material) AICc performs similarly to 10-fold CV for lasso, and that is what we will use for lasso. For other regularization methods, the selection rule will be 10-fold CV. According to our results (see Supplemental Material), SparseNet is slightly better than relaxed lasso and gamma lasso, and therefore we only present the results for SparseNet here.

A selected set of simulation results is presented in Table 5. Note that for BS, we only have results for $p \leq 30$, since it is fitted using the 'leaps' algorithm and $p \approx 30$ is the ad-hoc limit. We present the results in terms of % worse than the best possible BOSS, where the best possible BOSS means that on a single fit, we choose the subset size k_Q with the minimum RMSE among all $p + 1$ candidates, as if an oracle tells us the true model. Here is a brief summary of the results:

- For BOSS, AICc-hdf has a significant advantage over CV in terms of predictive performance, except for $n = 200$ and low SNR, in which case both selection rules are comparable. CV is also ten times heavier in terms of computation than AICc-hdf. These results are similar to the comparison of AICc-hdf and CV for BS with an orthogonal X as discussed in Section 2.3.2. Overall, the simulations indicate that AICc with hdf used in place of edf is a reasonable selection rule for an LS-based method that can be applied in practice without the requirement that the predictors are orthogonal. In the following discussions, when we refer to BOSS, we mean BOSS-AICc-hdf.

- The performance of BOSS is comparable to the performance of BS when BS is feasible. With a small sample size $n = 200$, BOSS performs either similarly to or better than BS for a high SNR, and it performs either similarly to or slightly worse than BS for a low SNR. With a large sample size $n = 2000$, BOSS is generally better than BS. Furthermore, BOSS only requires fitting the procedure once while BS uses CV as the selection rule, and a single fit of BOSS only has computational cost $O(np^2)$ so that BOSS is feasible for high dimensions.

- The performance of BOSS is generally better than the performance of FS. In the Dense model,

and Sparse-Ex3 with $n = 200$ and low SNR, we see that BOSS performs similarly to FS. In all other scenarios, the advantage of BOSS is obvious. For example, in Sparse-Ex4 with $n = 200$, high SNR and $\rho = 0.9$, FS is almost ten times worse than BOSS in terms of RMSE. Recall that Sparse-Ex4 is an example where FS has trouble stepping in all of the true predictors (with $\beta_j \neq 0$) in early steps. This is evidenced by the fact that FS chooses eight extra predictors on average in this situation, while BOSS only chooses approximately two extra predictors. Furthermore, FS based on CV is ten times computationally heavier than BOSS.

- Compared to the regularization methods, with a small sample size $n = 200$, BOSS is the best when SNR is high, lasso is the best when SNR is low and SparseNet is in between. With $n = 2000$, BOSS is almost always the best even when SNR is low. These findings are consistent with the discussion in Section 2.3.3, where we compare the performance of BS with regularization methods under an orthogonal X.

- In terms of support recovery in the sparse true models, LS-based methods can recover the true predictors (those with $\beta_j \neq 0$) and rarely include any noise predictors (those with $\beta_j = 0$) when SNR is high or the sample size n is large. However, SparseNet and lasso generally overfit, with the latter being worse in that regard. In the low SNR and small n scenario, lasso and SparseNet have more opportunity to recover the true predictors, but it comes with a price of including more false positives.

4 Real data analysis

In this section, we implement BOSS on five real world datasets. We consider four datasets from the StatLib library[2], which is maintained at Carnegie Mellon University. The 'Housing' data are often used in comparisons of different regression methods. The aim is to predict the housing values in the suburbs of Boston based on 13 predictors, including crime rate, property tax rate, pupil-teacher ratio, etc. The 'Hitters' data contain the 1987 annual salary for MLB players. For each player, it records 19 different performance metrics happening in 1986, such as number of times at bat, number of hits, etc., and the task is to predict the salary based on these statistics. The 'Auto' data are driven by prediction of the miles per gallon of vehicles based on features like the horsepower, weight, etc. The 'College' data contain various statistics for a large number of US colleges from the 1995 issue of 'US News and World Report', and we use these statistics to predict the number of applications received. We also consider a dataset from the Machine Learning Repository[3] that is maintained by UC Irvine. The 'ForestFire' data are provided by Cortez and Morais (2007) and the aim is to use meteorological and other data to predict the burned area of forest fires that happened in the northeast region of Portugal. The authors considered several machine learning algorithms, e.g. support vector regression, and concluded that the best prediction in terms of RMSE is the naive mean vector.

In real data analysis, one almost always would consider an intercept term. The way that BOSS handles the intercept term is described in Algorithm 2. To be more specific, we first center both X and y, and fit BOSS-AICc-hdf without an intercept to get $\hat{\beta}$. Then we calculate the intercept by $\hat{\beta}_0 = \bar{y} - \bar{X}^T\hat{\beta}$, which can be easily shown to be equivalent to fitting an intercept in every subset considered by BOSS.

We compare the performance of BOSS with LS-based methods BS and FS, and with regularization methods lasso and SparseNet. All of the methods are fitted with an intercept term. Note that for the Forest Fires dataset, we fit BS via MIO (Bertsimas et al., 2016) using the `R` package **bestsubset** (Hastie et al., 2017), where we restrict subset size $k = 0, \ldots, 10$, with 3 minutes as the time budget to find an

[2]http://lib.stat.cmu.edu/datasets/

[3]https://archive.ics.uci.edu/ml

Table 5: The performance of BOSS compared to other methods. Selection rules are for 'AICc-hdf/CV' for BOSS, AICc for lasso and CV for the remaining methods in the table, respectively.

				Sprse-Ex3					Sparse-Ex4					Dense				
				BOSS	BS	FS	lasso	SparseNet	BOSS	BS	FS	lasso	SparseNet	BOSS	BS	FS	lasso	SparseNet
				% worse than the best possible BOSS														
n=200	hsnr	$\rho = 0.5$	p=30	2/22	24	22	70	14	19/24	23	28	49	21	1/8	9	8	2	5
			p=180	1/18	-	21	135	17	4/15	-	16	82	13	14/13	-	16	47	8
		$\rho = 0.9$	p=30	5/41	17	41	66	12	21/33	21	56	73	28	2/9	10	8	2	8
			p=180	3/27	-	29	126	16	7/34	-	68	123	-10	15/12	-	12	71	20
	lsnr	$\rho = 0.5$	p=30	30/25	25	25	0	7	35/32	23	34	35	23	18/17	16	18	10	11
			p=180	11/13	-	13	-3	3	31/26	-	34	33	20	4/8	-	9	2	6
		$\rho = 0.9$	p=30	28/24	23	23	-2	5	32/27	18	78	71	44	15/14	15	13	10	12
			p=180	16/16	-	15	-5	1	17/18	-	36	33	35	12/10	-	10	15	9
n=2000	hsnr	$\rho = 0.5$	p=30	3/22	22	21	73	14	7/29	21	23	86	12	0/3	0	0	1	0
			p=180	1/22	-	22	130	14	6/28	-	21	174	12	8/9	-	12	37	8
		$\rho = 0.9$	p=30	2/21	21	22	74	13	32/33	16	33	108	12	0/3	1	1	2	1
			p=180	1/21	-	22	135	14	15/25	-	90	226	10	10/10	-	9	39	17
	lsnr	$\rho = 0.5$	p=30	5/22	21	21	73	14	13/30	22	23	61	15	2/9	10	10	2	7
			p=180	5/22	-	22	129	13	8/27	-	20	125	10	11/13	-	16	32	10
		$\rho = 0.9$	p=30	5/21	20	21	53	3	27/34	16	40	85	12	3/11	11	9	3	8
			p=180	4/17	-	17	92	-5	14/27	-	104	179	20	12/13	-	11	40	18
				Relative efficiency														
n=200	hsnr	$\rho = 0.5$	p=30	1/0.84	0.82	0.84	0.6	0.9	1/0.96	0.97	0.93	0.8	0.99	0.98/0.93	0.91	0.93	0.98	0.94
			p=180	1/0.85	-	0.84	0.43	0.86	1/0.91	-	0.9	0.57	0.92	0.95/0.96	-	0.93	0.73	1
		$\rho = 0.9$	p=30	1/0.74	0.9	0.74	0.63	0.93	1/0.91	1	0.78	0.7	0.95	0.98/0.91	0.9	0.92	0.98	0.93
			p=180	1/0.8	-	0.79	0.45	0.88	0.84/0.68	-	0.54	0.41	1	0.97/1	-	1	0.65	0.93
	lsnr	$\rho = 0.5$	p=30	0.77/0.8	0.8	0.8	1	0.93	0.91/0.93	1	0.92	0.91	1	0.93/0.94	0.95	0.93	1	0.99
			p=180	0.87/0.86	-	0.86	1	0.94	0.92/0.96	-	0.9	0.91	1	0.97/0.94	-	0.94	1	0.96
		$\rho = 0.9$	p=30	0.76/0.79	0.8	0.8	1	0.93	0.89/0.93	1	0.66	0.69	0.82	0.96/0.97	0.96	0.97	1	0.98
			p=180	0.82/0.82	-	0.83	1	0.95	1/0.99	-	0.86	0.88	0.87	0.97/0.99	-	1	0.95	1
n=2000	hsnr	$\rho = 0.5$	p=30	1/0.85	0.85	0.85	0.59	0.91	1/0.83	0.89	0.87	0.58	0.96	0.98/0.96	0.98	0.98	0.97	0.98
			p=180	1/0.83	-	0.83	0.44	0.89	1/0.83	-	0.88	0.39	0.95	1/0.99	-	0.96	0.79	1
		$\rho = 0.9$	p=30	1/0.84	0.85	0.84	0.59	0.9	0.84/0.84	0.96	0.84	0.54	1	1/0.97	0.99	0.99	0.98	0.99
			p=180	1/0.84	-	0.83	0.43	0.89	0.96/0.88	-	0.58	0.34	1	0.99/0.99	-	1	0.78	0.93
	lsnr	$\rho = 0.5$	p=30	1/0.86	0.86	0.86	0.61	0.92	1/0.87	0.93	0.92	0.7	0.99	0.98/0.91	0.9	0.91	0.98	0.93
			p=180	1/0.86	-	0.86	0.46	0.93	1/0.85	-	0.9	0.48	0.98	1/0.97	-	0.95	0.83	1
		$\rho = 0.9$	p=30	0.98/0.85	0.86	0.85	0.67	1	0.88/0.83	0.97	0.8	0.61	1	1/0.92	0.92	0.94	1	0.95
			p=180	0.91/0.81	-	0.81	0.49	1	1/0.9	-	0.56	0.41	0.95	0.99/0.99	-	1	0.8	0.94
				Sparsistency (number of extra variables)														
n=200	hsnr	$\rho = 0.5$	p=30	6(0)/6(0.6)	6(0.7)	6(0.6)	6(7.9)	6(1.1)	4.4(0.2)/5(1)	5(1)	4.8(1.1)	5.7(10.4)	4.8(2.1)	29.6/26.1	25.1	26	29.1	27
			p=180	6(0)/6(0.3)	-	6(0.4)	6(16.6)	6(2.4)	4(0)/4.2(0.5)	-	4.1(0.5)	5.1(20.2)	4.2(3.5)	17/20.2	-	19.6	52.2	32.4
		$\rho = 0.9$	p=30	6(0.6)/6(2.1)	6(0.8)	6(2.1)	6(9.2)	6(1.6)	5.1(2.8)/5.3(3.8)	5(1)	4.8(4.1)	5.8(17.8)	4.4(2.7)	29.3/25.2	23	24.6	28.9	26.2
			p=180	6(0.1)/6(0.6)	-	6(0.6)	6(16.2)	6(2.4)	4.2(2.4)/4.3(4.3)	-	4.3(8)	4.6(44.2)	4.1(3.1)	15.6/21.3	-	17.2	54.4	37.7
	lsnr	$\rho = 0.5$	p=30	2.9(2)/3.6(2.4)	3.4(2.1)	3.5(2.3)	5.1(6.9)	4.7(5.3)	2.3(1)/2.7(1.3)	2.6(1)	2.6(1.5)	3.6(6.9)	2.8(2.9)	5.7/7.5	6.6	7.1	5.3	10.3
			p=180	0.3(0.1)/1(0.7)	-	1(0.7)	3(9.7)	2.6(9.3)	1(0.2)/1.6(0.9)	-	1.3(0.8)	2.2(10.7)	2.1(6.5)	0.2/1.1	-	0.9	2.7	6.9
		$\rho = 0.9$	p=30	1.9(2.3)/2.4(3)	2.5(2.8)	2.4(2.9)	3.9(7.5)	3.7(6.1)	2.7(3.9)/3.2(5)	2.7(0.9)	2.2(4.4)	3(10.1)	2.9(8)	4.1/5.2	4.3	4.5	8.4	5.9
			p=180	0.5(0.2)/1.1(1.1)	-	1.1(1.1)	3.2(11.1)	3(10.8)	0.7(1.7)/1.1(5.5)	-	0.2(0.6)	0.3(4.8)	0.6(9)	1/2.1	-	1.6	3.9	3.3
n=2000	hsnr	$\rho = 0.5$	p=30	6(0.1)/6(0.6)	6(0.6)	6(0.6)	6(8.4)	6(1)	6(0.2)/6(0.6)	6(0.6)	6(0.6)	6(10.9)	6(0.8)	30/30	30	30	30	30
			p=180	6(0)/6(0.4)	-	6(0.4)	6(21.5)	6(2.3)	6(0.1)/6(0.3)	-	6(0.3)	6(32.2)	6(1.3)	34.5/35.1	-	32.6	106.5	43
		$\rho = 0.9$	p=30	6(0)/6(0.6)	6(0.6)	6(0.6)	6(9.2)	6(1)	6(0.4)/6(0.7)	6(0.6)	6(1.3)	6(17.8)	6(1.6)	30/29.9	29.9	29.9	30	30
			p=180	6(0)/6(0.4)	-	6(0.4)	6(23.2)	6(2.2)	6(1.4)/6(1.7)	-	5.9(3.8)	6(72.7)	6(8.7)	35/38.6	-	30.2	109.6	52.4
	lsnr	$\rho = 0.5$	p=30	6(0.1)/6(0.6)	6(0.6)	6(0.6)	6(8.3)	6(0.7)	4.2(0.4)/4.3(0.7)	4.3(0.6)	4.2(0.7)	5.2(9.6)	4.3(1)	29/22.7	21.2	22.1	28	24.1
			p=180	6(0.1)/6(0.4)	-	6(0.4)	6(21.2)	6(0.9)	4(0.1)/4(0.4)	-	4(0.4)	4.6(26.1)	4.1(1.3)	16/17	-	14.3	61.8	25.3
		$\rho = 0.9$	p=30	5.8(0.3)/5.8(1.1)	5.8(1.1)	5.8(1.1)	6(9.2)	6(0.8)	4.4(1.9)/4.4(1.7)	4.3(0.6)	4.3(2.2)	5.4(16.6)	4.2(2.5)	28.8/21.2	18.5	20.2	27.6	23.5
			p=180	5.7(0.3)/5.7(0.7)	-	5.7(0.7)	6(23)	6(1)	4.1(3.6)/4.1(4.4)	-	3.7(3.7)	4.6(60.3)	4.2(14.2)	16.6/21.4	-	11.8	65.3	32.3

optimal solution for each k, as suggested by the authors. For all of the other datasets, BS is fitted using the **leaps** package. To measure the performance of each method, we apply the leave-one-out (LOO) testing procedure, in which we fit the method on all observations except one, test the performance on that particular observation, and repeat the procedure for all n observations.

Table 6 presents the average RMSE, the average number of predictors and average running time for various methods given by LOO testing. We see that BOSS provides the best predictive performance in all datasets except the 'Housing' and 'College' data where lasso is the best for those datasets and its RMSE is 0.3% and 0.04% lower than those of BOSS, respectively. Due to a highly optimized implementation of the cyclical coordinate descent, the 'glmnet' algorithm is extremely fast to provide the lasso solution. BS is still not scalable to large dimensions, even by using the modern tools. With the dimension $p = 55$, it takes around 350 seconds to perform 10-fold CV for subset sizes restricted to be no greater than 10. However, We observe that BOSS is reasonably computationally efficient and much faster than BS, FS and SparseNet.

Table 6: Performance of subset selection methods on real datasets. The results are averages of leave-one-out (LOO) testing. The selection rules are AICc-hdf for BOSS, AICc for lasso and 10-fold CV for the rest, respectively. The intercept term is always fitted and is not counted in the number of predictors. Minimal values for the metrics for each dataset are given in bold face.

Dataset (n, p)	Metrics	BOSS	BS	FS	lasso	SparseNet
Housing (506, 13)	RMSE	3.372	3.37	3.383	**3.363**	3.369
	# predictors	12.004	**12.002**	12.026	12.012	**12.002**
	running time (s)	0.021	0.066	0.156	**0.007**	0.39
Hitters (263, 19)	RMSE	**233.853**	236.989	238.222	234.064	238.375
	# predictors	11.152	10.852	**10.662**	14.205	12.51
	running time (s)	0.014	0.095	0.104	**0.008**	0.493
Auto (392, 6)	RMSE	**2.628**	**2.628**	**2.628**	2.643	2.63
	# predictors	**3**	3.003	**3**	5.008	3.008
	running time (s)	0.008	0.051	0.067	**0.007**	0.25
College (777, 17)	RMSE	1565.476	1568.234	1569.625	**1564.807**	1573.975
	# predictors	17.991	16.333	16.067	16.008	**15.385**
	running time (s)	0.058	0.092	0.451	**0.01**	0.734
Forest Fires (517, 55)	RMSE	**18.603**	18.707	18.757	18.726	19.163
	# predictors	**0**	0.983	0.986	2.985	6.899
	running time (s)	0.084	356.651	0.593	**0.014**	0.785

5 Conclusion and future work

In this paper, we introduce a heuristic degrees of freedom (hdf) for BS based on an orthogonal X. We further propose a KL-based information criterion AICc-edf and its feasible implementation AICc-hdf. We show that their expected values can reasonably approximate the expected KL, $E(\mathrm{Err}_{\mathrm{KL}})$. Moreover, they result in the same choice of subset as $\widehat{\mathrm{Err}}_{\mathrm{KL}}$ when they are used as selection rules for BS. Furthermore, we propose an LS-based subset selection method BOSS. BOSS together with the selection rule AICc-hdf has computational cost on the same order as OLS. Finally, we show in simulations and real data examples that BOSS can be a competitive method in both speed and predictive performance.

Since edf (3) for LS-based methods is saturated at n when $p \geq n$, potential future work is to study a measure of complexity and build a connection with the use of information criteria in the case of $p \geq n$. The strong performance of BOSS using AICc compared to using CV suggests that the pursuit of methods to approximate edf (which normally does not have an analytical expression for complex modeling methods and algorithms), particularly for methods that are more sensitive to small perturbations in the data, is worthy of further research.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. P. F. Csaki (Ed.), *2nd International Symposium on Information Theory*, Budapest, Hungary, pp. 267–281. Akadmiai Kiad.

Bertsimas, D., A. King, and R. Mazumder (2016). Best subset selection via a modern optimization lens. *The Annals of Statistics 44*(2), 813–852.

Cortez, P. and A. Morais (2007). A data mining approach to predict forest fires using meteorological data. In J. Neves, M. F. Santos and J. Machado Eds. *New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence*, 512–523.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association 81*(394), 461–470.

Efron, B. (2004). The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association 99*(467), 619–632.

Efroymson, M. (1960). Multiple regression analysis. In A. Ralston and H. S. Wilf, Eds. *Mathematical Methods for Digital Computers*, 191–203.

Flynn, C. J., C. M. Hurvich, and J. S. Simonoff (2013). Efficiency for regularization parameter selection in penalized likelihood estimation of misspecified models. *Journal of the American Statistical Association 108*(503), 1031–1043.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1.

Furnival, G. M. and R. W. Wilson (1974). Regressions by leaps and bounds. *Technometrics 16*(4), 499–511.

Hammarling, S. and C. Lucas (2008). Updating the qr factorization and the least squares problem. MIMS EPrint: 2008.111, reports available from http://eprints.maths.manchester.ac.uk/.

Harris, X. T. (2016). Prediction error after model search. *arXiv preprint arXiv:1610.06107*.

Hastie, T., R. Tibshirani, and R. J. Tibshirani (2017). Extended comparisons of best subset selection, forward stepwise selection, and the lasso. *arXiv preprint arXiv:1707.08692*.

Hocking, R. and R. Leslie (1967). Selection of the best subset in regression analysis. *Technometrics 9*(4), 531–540.

Hurvich, C. M. and C.-L. Tsai (1989). Regression and time series model selection in small samples. *Biometrika 76*(2), 297–307.

Hurvich, C. M. and C.-L. Tsai (1991). Bias of the corrected AIC criterion for underfitted regression and time series models. *Biometrika 78*(3), 499–509.

Janson, L., W. Fithian, and T. J. Hastie (2015). Effective degrees of freedom: a flawed metaphor. *Biometrika 102*(2), 479–485.

Konishi, S. and G. Kitagawa (2008). *Information Criteria and Statistical Modeling*. Berlin: Springer Science & Business Media.

Liao, J., J. E. Cavanaugh, and T. L. McMurry (2018). Extending AIC to best subset regression. *Computational Statistics 33*(2), 787–806.

Mallows, C. L. (1973). Some comments on Cp. *Technometrics 15*(4), 661–675.

Mazumder, R., J. H. Friedman, and T. Hastie (2011). Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association 106*(495), 1125–1138.

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics & Data Analysis 52*(1), 374–393.

Meinshausen, N. (2012). *relaxo: Relaxed Lasso*. R package version 0.1-2.

Natarajan, B. K. (1995). Sparse approximate solutions to linear systems. *SIAM Journal on Computing 24*(2), 227–234.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*(2), 461–464.

Taddy, M. (2017). One-step estimator paths for concave regularization. *Journal of Computational and Graphical Statistics 26*(3), 525–536.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*(1), 267–288.

Tibshirani, R. J. (2015). Degrees of freedom and model search. *Statistica Sinica*, 1265–1296.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association 93*(441), 120–131.

Supplementary Materials
On the Use of Information Criteria for Subset Selection in Least Squares Regression

Sen Tian, Clifford M. Hurvich, Jeffrey S. Simonoff

A Technical details

A.1 Proof of theorem 1 and its corollary

In this section, we assume an orthogonal X and a null true model. This is the only scenario under which both $\mathrm{df}_C(k)$ and $\mathrm{hdf}(k)$ have analytical expressions. We will prove that the ratio of $\mathrm{df}_C(k)$ and $\mathrm{hdf}(k)$ goes to 1 as $k, p \to \infty$ while $k = \lfloor xp \rfloor$, where $\lfloor \cdot \rfloor$ denotes the greatest integer function and $x \in (0, 1)$. We start by laying out a few lemmas to be used in the proof of the main theorem.

Lemma 1. *Assume the design matrix is orthogonal and the true model is null ($\mu = 0$). Then*

$$hdf(k) = df_L(\lambda_k^\star) = k - 2p \cdot \Phi^{-1}\left(\frac{k}{2p}\right) \cdot \phi\left[\Phi^{-1}\left(\frac{k}{2p}\right)\right]. \tag{19}$$

Proof. We follow the steps described in algorithm 1. We first find $\lambda_k^\star$ from (10), by using the fact that $\mu = 0$, and we get $-\dfrac{\sqrt{2\lambda_k^\star}}{\sigma} = \Phi^{-1}\left(\dfrac{k}{2p}\right)$, which we then substituted into (9) to get (19). □

Lemma 2. *Define $\tilde{G}(x) = x - \Phi^{-1}(x) \cdot \phi\left[\Phi^{-1}(x)\right]$, where $x \in (0, 1)$ is a continuous variable. We have*

$$\lim_{x \to 0} \tilde{G}(x) = 0,$$

and

$$\tilde{G}'(x) = \left[\Phi^{-1}(x)\right]^2.$$

Therefore by the fundamental theorem of calculus,

$$\tilde{G}(x) = \int_0^x \left[\Phi^{-1}(u)\right]^2 du.$$

Proof. First note that, since $\phi'(v) = -v \cdot \phi(v)$ and $\lim_{v \to \pm\infty} \phi'(v) = 0$, we have

$$\lim_{v \to \pm\infty} v \cdot \phi(v) = 0.$$

Let $v = \Phi^{-1}(x)$. Then

$$\lim_{x \to 0} \tilde{G}(x) = \lim_{v \to -\infty} -v \cdot \phi(v) = 0.$$

Next, we obtain the derivative of $\tilde{G}(x)$. Since $\Phi'(x) = \phi(x)$, we have

$$\left[\Phi^{-1}(x)\right]' = \frac{1}{\Phi'\left[\Phi^{-1}(x)\right]} = \frac{1}{\phi\left[\Phi^{-1}(x)\right]}. \tag{20}$$

Also since $\phi'(x) = -x \cdot \phi(x)$, we have

$$\phi'\left[\Phi^{-1}(x)\right] = -\Phi^{-1}(x) \cdot \phi\left[\Phi^{-1}(x)\right] \cdot \left[\Phi^{-1}(x)\right]' = -\Phi^{-1}(x). \tag{21}$$

By (20) and (21), we have

$$\tilde{G}'(x) = 1 - \left[\Phi^{-1}(x)\right]' \cdot \phi\left[\Phi^{-1}(x)\right] - \left[\Phi^{-1}(x)\right] \cdot \phi'\left[\Phi^{-1}(x)\right] = \left[\Phi^{-1}(x)\right]^2.$$

Therefore, by the fundamental theorem of calculus, we have

$$\tilde{G}(x) = \int_0^x \tilde{G}'(u)du + \tilde{G}(0) = \int_0^x \left[\Phi^{-1}(u)\right]^2 du.$$

□

Lemma 3. *Denote $\tilde{Q}$ as the quantile function of a χ_1^2 distribution, and let $\tilde{H}(s) = -\tilde{Q}(1-s)$ where $s \in (0,1)$. For $0 \le s \le t \le 1$, consider the truncated variance function*

$$\tilde{\sigma}^2(s,t) = \int_s^t \int_s^t (u \wedge v - uv) d\tilde{H}(u) d\tilde{H}(v), \tag{22}$$

where $u \wedge v = \min(u,v)$. We have

$$0 \le \tilde{\sigma}^2(s,t) \le 1.$$

Proof. We first note three facts.

$$\tilde{H}(s) = -\left[\Phi^{-1}\left(1 - \frac{s}{2}\right)\right]^2 = -\left[\Phi^{-1}\left(\frac{s}{2}\right)\right]^2, \tag{23}$$

$$d\tilde{H}(s) = \frac{\Phi^{-1}(1-s/2)}{\phi\left[\Phi^{-1}(1-s/2)\right]} ds = -\frac{\Phi^{-1}(s/2)}{\phi\left[\Phi^{-1}(s/2)\right]} ds, \quad \text{by (20)}, \tag{24}$$

$$\Phi^{-1}(w) = -\sqrt{\log\frac{1}{w^2} - \log\log\frac{1}{w^2} - \log(2\pi)} + o(1), \quad \text{for small } w, \text{ by Fung and Seneta (2017)}. \tag{25}$$

Hence for small w,

$$\left[\Phi^{-1}(w)\right]^2 = O\left(\log\frac{1}{w^2}\right). \tag{26}$$

Then by (23) and (26), we have

$$\lim_{s\to 0} s \cdot \tilde{H}(s) = \lim_{s\to 0} -s \cdot \left[\Phi^{-1}\left(\frac{s}{2}\right)\right]^2 = 0. \tag{27}$$

Also, by (23) and Lemma 2,

$$-\int_0^x \tilde{H}(s)ds = 2 \cdot \tilde{G}\left(\frac{x}{2}\right). \tag{28}$$

Since $u, v \in [0,1]$, we have $u \wedge v - uv \ge 0$. By (24), we have $d\tilde{H}(s)/ds \ge 0$. Therefore, the integrand in (22) is non-negative, so that

$$\tilde{\sigma}^2(s,t) \ge 0,$$

and

$$\begin{aligned}
\tilde{\sigma}^2(s,t) &\le \int_0^1 \int_0^1 (u \wedge v - uv) d\tilde{H}(u) d\tilde{H}(v), \\
&= \int_0^1 \left[\int_0^v u(1-v)d\tilde{H}(u) + \int_v^1 v(1-u)d\tilde{H}(u)\right] d\tilde{H}(v), \\
&= \int_0^1 \left[\int_0^v u d\tilde{H}(u) + v\int_v^1 d\tilde{H}(u) - v\int_0^1 u d\tilde{H}(u)\right] d\tilde{H}(v).
\end{aligned}$$

Denote

$$\tilde{M}(v) = \int_0^v u d\tilde{H}(u) + v\int_v^1 d\tilde{H}(u) - v\int_0^1 u d\tilde{H}(u).$$

Now, we consider the three integrals in $\tilde{M}(v)$. First note that

$$\begin{aligned}\int_0^x u d\tilde{H}(u) &= u \cdot \tilde{H}(u)\Big|_0^x - \int_0^x \tilde{H}(u)du, \\ &= x \cdot \tilde{H}(x) - \int_0^x \tilde{H}(u)du, \quad \text{by (27)} \\ &= x \cdot \tilde{H}(x) + 2 \cdot \tilde{G}(x/2), \quad \text{by (28)}.\end{aligned}$$

It is easily verified that $\tilde{H}(1) = 0$ and $\tilde{G}(1/2) = 1/2$, we have

$$\int_0^1 u d\tilde{H}(u) = 2 \cdot \tilde{G}(1/2) = 1,$$

and

$$v \int_v^1 d\tilde{H}(u) = -v \cdot \tilde{H}(v).$$

Therefore,

$$\begin{aligned}\tilde{M}(v) &= v \cdot \tilde{H}(v) + 2 \cdot \tilde{G}(v/2) - v \cdot \tilde{H}(v) - 2v \cdot \tilde{G}(1/2) \\ &= 2 \cdot \tilde{G}(v/2) - v.\end{aligned}$$

Finally,

$$\begin{aligned}\int_0^1 \tilde{M}(v) d\tilde{H}(v) &= \int_0^1 2 \cdot \tilde{G}(v/2) d\tilde{H}(v) - \int_0^1 v d\tilde{H}(v), \\ &= -\int_0^1 \Phi^{-1}\left(\frac{v}{2}\right) \cdot \phi\left[\Phi^{-1}\left(\frac{v}{2}\right)\right] d\tilde{H}(v), \quad \text{by the definition of } \tilde{G}(x), \\ &= 2\int_0^{1/2} \left[\Phi^{-1}(v)\right]^2 dv, \quad \text{by (24)}, \\ &= 2 \cdot \tilde{G}(1/2), \\ &= 1.\end{aligned}$$

Therefore,

$$0 \le \tilde{\sigma}^2(s,t) \le 1.$$

□

Theorem 3. *Assume the design matrix is orthogonal and the true model is null ($\mu = 0$). Let $\tilde{X}_{(i)}$ be the i-th largest order statistic in an i.i.d sample of size p from a χ_1^2 distribution. Denote $\tilde{Y}_p = \tilde{\sigma}_p^{-1}(\sum_{i=1}^k \tilde{X}_{(i)} - \tilde{\mu}_p)$, where*

$$\tilde{\sigma}_p = \sqrt{p} \cdot \sigma(1/p, k/p),$$

and

$$\tilde{\mu}_p = -p\int_{1/p}^{k/p} \tilde{H}(u)du - \tilde{H}\left(\frac{1}{p}\right),$$

where $\sigma(s,t)$ and $\tilde{H}(x)$ are defined in Lemma 3.

As $k \to \infty$, $p \to \infty$ and $k = \lfloor px \rfloor$ with $x \in (0,1)$, we have

$$\frac{df_C(k)}{2p} = \frac{1}{2p} E\left[\sum_{i=1}^k \tilde{X}_{(i)}\right] = \frac{\tilde{\sigma}_p}{2p} E(\tilde{Y}_p) + \tilde{G}\left(\frac{k}{2p}\right) + O\left(\frac{\log(p)}{p}\right), \tag{29}$$

where $\lfloor \cdot \rfloor$ denotes the greatest integer function, $\tilde{G}(x)$ is defined in Lemma 2.

Proof. We first apply a result in Csorgo et al. (1991), to show that $\tilde{Y}_p = \tilde{\sigma}_p^{-1}(\sum_{i=1}^{k} \tilde{X}_{(i)} - \tilde{\mu}_p)$ converges in distribution to a standard normal. We then show how $\tilde{\mu}_p$ can be expressed in terms of function G plus a remainder term, which further leads to expression (29).

It follows from Csorgo et al. (1991) Corollary 2, that if there exist centering and normalizing constants c_p and $d_p > 0$, s.t.

$$d_p^{-1}(\tilde{X}_{(1)} - c_p) \xrightarrow{D} Y, \quad \text{where Y is the standard Gumbel distribution,} \tag{30}$$

then as $k \to \infty$, $p \to \infty$ and $k = \lfloor px \rfloor$ with $x \in (0, 1)$,

$$\left(\sum_{i=1}^{k} \tilde{X}_{(i)} - \tilde{\mu}_p \right) / \tilde{\sigma}_p \xrightarrow{D} Z, \quad \text{where Z is standard normal.} \tag{31}$$

First, it follows from Embrechts et al. (2013) that (30) holds, with $c_p = 2\log(p) - \log\log(p) - \log(\pi)$ and $d_p = 2$.

Next, we have

$$\begin{aligned}
\tilde{\mu}_p &= -p \int_{1/p}^{k/p} \tilde{H}(u)du - \tilde{H}\left(\frac{1}{p}\right), \\
&= -p \int_{0}^{k/p} \tilde{H}(u)du + p \int_{0}^{1/p} \tilde{H}(u)du - \tilde{H}\left(\frac{1}{p}\right), \\
&= 2p \cdot \tilde{G}\left(\frac{k}{2p}\right) - 2p \cdot \tilde{G}\left(\frac{1}{2p}\right) + \left[\Phi^{-1}\left(\frac{1}{2p}\right)\right]^2, \quad \text{by (28).}
\end{aligned}$$

Also, since

$$\begin{aligned}
\tilde{G}(\frac{1}{2p}) &= \frac{1}{2p} - \Phi^{-1}\left(\frac{1}{2p}\right) \cdot \phi\left[\Phi^{-1}\left(\frac{1}{2p}\right)\right], \quad \text{by definition of } \tilde{G}(x) \text{ in Lemma 2,} \\
&= \frac{1}{2p} - \frac{1}{\sqrt{2\pi}} \Phi^{-1}\left(\frac{1}{2p}\right) \cdot \exp\left(-\frac{1}{2}\left[\Phi^{-1}\left(\frac{1}{2p}\right)\right]^2\right), \\
&= \frac{1}{2p} + \frac{1}{\sqrt{2\pi}} \cdot \left(\sqrt{\log(4p^2) - \log\log(4p^2) - \log(2\pi)} + o(1)\right) \cdot \\
&\qquad \exp\left[-\frac{1}{2}\left(\log(4p^2) - \log\log(4p^2) - \log(2\pi) + o(1)\right)\right], \quad \text{by (25),} \\
&= \frac{1}{2p} + \left(\sqrt{\log(4p^2) - \log\log(4p^2) - \log(2\pi)} + o(1)\right) \cdot \frac{\sqrt{\log(4p^2)}}{2p}, \\
&= O\left(\frac{\log(p)}{p}\right).
\end{aligned}$$

Also

$$\frac{1}{2p}\left[\Phi^{-1}\left(\frac{1}{2p}\right)\right]^2 = O\left(\frac{\log(p)}{p}\right), \quad \text{by (25),}$$

and hence

$$\begin{aligned}
\frac{\tilde{\mu}_p}{2p} &= \tilde{G}\left(\frac{k}{2p}\right) - \tilde{G}\left(\frac{1}{2p}\right) + \frac{1}{2p}\left[\Phi^{-1}\left(\frac{1}{2p}\right)\right]^2, \\
&= \tilde{G}\left(\frac{k}{2p}\right) + O\left(\frac{\log(p)}{p}\right).
\end{aligned}$$

Therefore, (29) holds, i.e.

$$\frac{\mathrm{df}_C(k)}{2p} = \frac{1}{2p} E\left(\sum_{i=1}^{k} \tilde{X}_{(i)}\right) = \frac{\tilde{\sigma}_p}{2p} E(\tilde{Y}_p) + \frac{\tilde{\mu}_p}{2p} = \frac{\tilde{\sigma}_p}{2p} E(\tilde{Y}_p) + \tilde{G}\left(\frac{k}{2p}\right) + O\left(\frac{\log(p)}{p}\right).$$

□

Corollary 3.1. *If* $\limsup |E(\tilde{Y}_p)| < \infty$*, we further have:*

$$\frac{df_C(k)}{2p} = \tilde{G}\left(\frac{k}{2p}\right) + O\left(\frac{\log(p)}{p}\right) + O\left(\frac{1}{\sqrt{p}}\right). \tag{32}$$

Proof. By Lemma 3 we have $0 \le \sigma(1/p, k/p) \le 1$, and hence $\tilde{\sigma}_p = O(\sqrt{p})$. Therefore by Theorem 3, we have

$$\frac{\mathrm{df}_C(k)}{2p} = \tilde{G}\left(\frac{k}{2p}\right) + O\left(\frac{\log(p)}{p}\right) + O\left(\frac{1}{\sqrt{p}}\right).$$

□

Theorem 1. *Assume* X *is orthogonal and the true model is null* ($\mu = 0$)*. As* $p \to \infty$*,* $k \to \infty$ *with* $k = \lfloor px \rfloor$*, we have*

$$\frac{1}{2p} hdf(k) = \frac{1}{2p} df_C(k) - \frac{\tilde{\sigma}_p}{2p} E(\tilde{Y}_p) + O\left(\frac{\log(p)}{p}\right), \tag{11}$$

where $x \in (0, 1)$ *is a constant and* $\lfloor \cdot \rfloor$ *denotes the greatest integer function.*

Proof. By Lemma 1, we have

$$\mathrm{hdf}(k) = df_L(\tilde{M}^{-1}(k)) = k - 2p \cdot \Phi^{-1}\left(\frac{k}{2p}\right) \cdot \phi\left[\Phi^{-1}\left(\frac{k}{2p}\right)\right].$$

Then by the definition of $\tilde{G}(x)$ in Lemma 2,

$$\frac{1}{2p}\mathrm{hdf}(k) = \tilde{G}\left(\frac{k}{2p}\right).$$

By Theorem 3, we also have

$$\frac{1}{2p}\mathrm{df}_C(k) = \frac{\sigma_p}{2p} E(\tilde{Y}_p) + \tilde{G}\left(\frac{k}{2p}\right) + O\left(\frac{\log(p)}{p}\right).$$

Therefore, (11) holds, i.e.

$$\frac{1}{2p}\mathrm{hdf}(k) = \frac{1}{2p}\mathrm{df}_C(k) - \frac{\tilde{\sigma}_p}{2p} E(\tilde{Y}_p) + O\left(\frac{\log(p)}{p}\right).$$

□

Corollary 1.1. *If* $\limsup |E(\tilde{Y}_p)| < \infty$ *, we further have*

$$\frac{df_C(k)}{hdf(k)} \to 1. \tag{12}$$

Proof. By Theorem 1 and Corollary 3.1,

$$\frac{1}{2p}\mathrm{hdf}(k) = \frac{1}{2p}\mathrm{df}_C(k) + O\left(\frac{1}{\sqrt{p}}\right) + O\left(\frac{\log(p)}{p}\right).$$

From Lemma 2, $\tilde{G}(x)$ is a non-decreasing function with $\tilde{G}(0+) = 0$ and $\tilde{G}(1/2) = 1/2$. Thus,

$$\frac{2p}{\mathrm{hdf}(k)} = \frac{1}{\tilde{G}\left(\frac{k}{2p}\right)} = O(1),$$

since $k = \lfloor px \rfloor$ and $x \in (0,1)$. Therefore,

$$\frac{\mathrm{df}_C(k)}{\mathrm{hdf}(k)} = 1 + O\left(\frac{1}{\sqrt{p}}\right) + O\left(\frac{\log(p)}{p}\right),$$

and hence

$$\frac{\mathrm{df}_C(k)}{\mathrm{hdf}(k)} \to 1.$$

□

A.2 Expected KL-based optimism, in the context of BS

In this section, we obtain the expected Kullback-Leibler (KL) based optimism for BS with subset size k. Let's first consider fitting least squares regression on k prefixed predictors. Recall that

$$y = \mu + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. We use the deviance to measure the predictive error, that is

$$\Theta = -2\log f(y|\mu, \sigma^2).$$

The training error is then

$$\mathrm{err}_{\mathrm{KL}} = -2\log f(y|\hat{\mu}, \hat{\sigma}^2),$$

and the testing error (KL information) is

$$\mathrm{Err}_{\mathrm{KL}} = -2E_0\left[\log f(y^0|\hat{\mu}, \hat{\sigma}^2)\right],$$

where $\hat{\mu}$ and $\hat{\sigma}^2$ are the maximum likelihood estimators (MLE) based on training data (X, y), y^0 is independent and has the same distribution of y and E_0 is the expectation over y^0.

Due to the assumption of normality, the deviance can be expressed as

$$\Theta = n\log(2\pi\sigma^2) + \frac{\|y-\mu\|_2^2}{\sigma^2}. \tag{33}$$

Maximizing the likelihood, or minimizing the deviance (33), gives

$$\begin{aligned} \hat{\mu} &= \arg\min_{\mu} \|y-\mu\|_2^2, \\ \hat{\sigma}^2 &= \frac{1}{n}\|y-\hat{\mu}\|_2^2. \end{aligned} \tag{34}$$

Using these expressions, we then have

$$\mathrm{err}_{\mathrm{KL}} = n\log(2\pi\hat{\sigma}^2) + n, \tag{35}$$

and

$$\mathrm{Err}_{\mathrm{KL}} = n\log(2\pi\hat{\sigma}^2) + n\frac{\sigma^2}{\hat{\sigma}^2} + \frac{\|\mu - \hat{\mu}\|_2^2}{\hat{\sigma}^2}.$$

The expected optimism is then

$$\begin{aligned} E(\mathrm{op}_{\mathrm{KL}}) &= E(\mathrm{Err}_{\mathrm{KL}}) - E(\mathrm{err}_{\mathrm{KL}}), \\ &= E\left(n\frac{\sigma^2}{\hat{\sigma}^2}\right) + E\left(\frac{\|\mu - \hat{\mu})\|_2^2}{\hat{\sigma}^2}\right) - n. \end{aligned} \tag{36}$$

So far we've been considering a subset with k fixed predictors. At subset size k, BS chooses the one with minimum residual sum of squares (RSS) from all $\binom{p}{k}$ possible subsets. In order for the above derivation to continue to hold for BS of subset size k, we need to show that the MLE from (34) is also the BS fit. This can be easily obtained from the full likelihood (-2 times) (35), which after substituting the expression of $\hat{\sigma}$ leads to

$$n\log\left(\frac{2\pi}{n}\|y - \hat{\mu}\|_2^2\right) + n.$$

Therefore, for all $\binom{p}{k}$ models of size k, the one with largest log likelihood, is also the one with smallest RSS. Hence (36) holds for BS fit with subset size k as well.

A.3 Proof of Theorem 2

Proof. Since $[X_1, X_2, \cdots, X_j]$ and $[Q_1, Q_2, \cdots, Q_j]$ span the same space, we have

$$\hat{\alpha}^{(j)} = \hat{\beta}^{(j)}. \tag{37}$$

We can express $\hat{\gamma}(k_Q)$ as

$$\hat{\gamma}(k_Q) = \sum_{j \in S_k} \hat{\gamma}^{(j)} - \hat{\gamma}^{(j-1)}. \tag{38}$$

We multiply both sides by R^{-1} (X is assumed to have full column rank), and use (37) to get

$$\hat{\beta}(k_Q) = \sum_{j \in S_k} \hat{\alpha}^{(j)} - \hat{\alpha}^{(j-1)}.$$

□

B Complete simulation results

- Orthogonal X, simulation setups are discussed in Section 2.3.1.
  - The performance of selection rules for BS. The selection rules include C_p, AICc, BIC, GCV and 10-fold CV. For each selection rule except CV, there are two columns in the table indicating the degrees of freedoms to use in calculating the information criterion. The 'edf' (effective degrees of freedom) is estimated using definition (3) by assuming the knowledge of μ and σ, and hence it is an infeasible rule. The 'ndf/hdf/bdf' (naive degrees of freedom / heuristic degrees of freedom / degrees of freedom based on bootstrap) are feasible selection rules in practice.
    - Orth-Sparse-Ex1: tables S1-S2
    - Orth-Sparse-Ex2: tables S3-S4
    - Orth-Dense: tables S5-S6
  - The performance of BS and regularization methods. Note that for lasso, we use the number of non-zero coefficients $k(\lambda)$ in place of edf in the AICc formula (7). Zou et al. (2007) showed that $k(\lambda)$ is an unbiased estimator of edf for lasso. For gamma lasso, Taddy (2017) suggested a heuristic degrees of freedom to be plugged into (7) in order to use AICc as the selection rule.
    - Orth-Sparse-Ex1: tables S7-S8
    - Orth-Sparse-Ex2: tables S9-S10
    - Orth-Dense: tables S11-S12
- General X, simulation setups are discussed in Section 3.3.1.
  - The performance of BOSS, BS, FS, lasso, gamma lasso, SparseNet and relaxed lasso (rlasso).
    - Sparse-Ex1: tables S13-S18
    - Sparse-Ex2: tables S19-S24
    - Sparse-Ex3: tables S25-S30
    - Sparse-Ex4: tables S31-S36
    - Dense: tables S37-S42

Table S1: The performance of BS using different selection rules, Orth-Sparse-Ex1, n=200

		C_p		AICc		BIC		GCV		CV
		edf	ndf/hdf/bdf	edf	ndf/hdf/bdf	edf	ndf/hdf/bdf	edf	ndf/hdf/bdf	
		% worse than the best possible BS								
hsnr	p=14	8	33/9/11	7	33/6/9	0	10/0/1	9	34/8/10	19
	p=30	4	84/5/7	2	83/2/5	0	28/0/0	4	86/4/7	24
	p=60	2	157/3/5	1	159/2/3	0	64/0/0	2	167/3/4	-
	p=180	1	338/30/32	0	392/1/2	0	206/0/0	0	431/2/3	-
msnr	p=14	8	33/14/12	7	33/11/10	0	10/2/1	9	34/14/12	19
	p=30	4	84/12/11	2	83/8/7	0	28/1/2	4	86/10/10	24
	p=60	2	157/13/10	1	159/8/7	0	64/2/2	2	167/11/9	-
	p=180	1	338/40/38	0	392/7/6	0	206/3/4	0	431/10/8	-
lsnr	p=14	18	16/23/23	18	17/24/24	93	43/97/93	18	16/23/23	26
	p=30	20	25/36/33	21	24/37/35	68	23/68/67	21	26/37/35	28
	p=60	18	44/28/27	21	45/31/30	43	17/43/43	20	48/30/29	-
	p=180	15	108/35/34	18	132/22/22	25	50/25/25	17	149/22/21	-
		Relative efficiency								
hsnr	p=14	0.93	0.75/0.92/0.9	0.94	0.75/0.94/0.92	1	0.91/1/0.99	0.92	0.75/0.92/0.91	0.84
	p=30	0.96	0.54/0.95/0.93	0.98	0.55/0.98/0.96	1	0.78/1/1	0.96	0.54/0.96/0.94	0.81
	p=60	0.98	0.39/0.97/0.95	0.99	0.39/0.98/0.97	1	0.61/1/1	0.98	0.38/0.97/0.96	-
	p=180	0.99	0.23/0.77/0.76	1	0.2/0.99/0.98	1	0.33/1/1	1	0.19/0.98/0.97	-
msnr	p=14	0.93	0.75/0.88/0.89	0.94	0.75/0.9/0.91	1	0.91/0.99/0.99	0.92	0.75/0.88/0.9	0.84
	p=30	0.96	0.54/0.89/0.9	0.98	0.55/0.92/0.93	1	0.78/0.99/0.98	0.96	0.54/0.91/0.91	0.81
	p=60	0.98	0.39/0.89/0.91	0.99	0.39/0.92/0.93	1	0.61/0.98/0.98	0.98	0.38/0.9/0.92	-
	p=180	0.99	0.23/0.71/0.72	1	0.2/0.93/0.94	1	0.33/0.97/0.96	1	0.19/0.91/0.92	-
lsnr	p=14	0.99	1/0.95/0.95	0.98	1/0.94/0.94	0.6	0.82/0.59/0.6	0.99	1/0.95/0.94	0.92
	p=30	1	0.97/0.89/0.9	1	0.97/0.88/0.89	0.72	0.98/0.72/0.72	0.99	0.96/0.88/0.89	0.94
	p=60	0.99	0.81/0.92/0.92	0.97	0.81/0.89/0.9	0.82	1/0.82/0.82	0.98	0.79/0.9/0.91	-
	p=180	1	0.55/0.86/0.86	0.97	0.5/0.95/0.95	0.93	0.77/0.93/0.93	0.98	0.46/0.95/0.95	-
		Sparsistency (number of extra variables)								
hsnr	p=14	6(0.2)	6(1.3)/6(0.4)/6(0.4)	6(0.2)	6(1.2)/6(0.2)/6(0.3)	6(0)	6(0.2)/6(0)/6(0)	6(0.3)	6(1.3)/6(0.4)/6(0.4)	6(0.7)
	p=30	6(0.1)	6(3.9)/6(0.2)/6(0.2)	6(0)	6(3.8)/6(0.1)/6(0.1)	6(0)	6(0.6)/6(0)/6(0)	6(0.1)	6(4.1)/6(0.1)/6(0.1)	6(0.7)
	p=60	6(0)	6(8.9)/6(0)/6(0.1)	6(0)	6(9.2)/6(0)/6(0)	6(0)	6(1.6)/6(0)/6(0)	6(0)	6(10.5)/6(0)/6(0.1)	-
	p=180	6(0)	6(32.2)/6(6.4)/6(6.3)	6(0)	6(48.9)/6(0)/6(0)	6(0)	6(9.5)/6(0)/6(0)	6(0)	6(74.6)/6(0)/6(0)	-
msnr	p=14	6(0.2)	6(1.3)/6(0.7)/6(0.4)	6(0.2)	6(1.2)/6(0.5)/6(0.3)	6(0)	6(0.2)/6(0)/6(0)	6(0.3)	6(1.3)/6(0.6)/6(0.4)	6(0.7)
	p=30	6(0.1)	6(3.9)/6(0.4)/6(0.3)	6(0)	6(3.8)/6(0.2)/6(0.1)	6(0)	6(0.6)/6(0)/6(0)	6(0.1)	6(4.1)/6(0.3)/6(0.2)	6(0.7)
	p=60	6(0)	6(8.9)/6(0.3)/6(0.2)	6(0)	6(9.2)/6(0.1)/6(0.1)	6(0)	6(1.6)/6(0)/6(0)	6(0)	6(10.5)/6(0.2)/6(0.1)	-
	p=180	6(0)	6(32.2)/6(6.6)/6(6.6)	6(0)	6(48.9)/6(0.1)/6(0.1)	6(0)	6(9.5)/6(0)/6(0)	6(0)	6(74.6)/6(0.1)/6(0.1)	-
lsnr	p=14	5.5(2.3)	5.2(1.3)/5.6(4.6)/5.4(3.6)	5.4(2.1)	5.2(1.2)/5.4(4.2)/5.3(3.2)	0.9(0.1)	3.6(0.2)/0.7(0.1)/0.9(0.1)	5.5(2.4)	5.3(1.3)/5.6(4.6)/5.4(3.5)	4.9(1.6)
	p=30	4.5(1.9)	5.3(3.9)/4.2(4.9)/4.2(4)	4.2(1.2)	5.2(3.8)/3.3(2.2)/3.4(1.8)	0.1(0)	3.7(0.6)/0.1(0)/0.2(0)	4.5(2)	5.3(4.1)/3.9(4.1)/3.9(3.3)	4(1.9)
	p=60	3.4(1.1)	5.2(8.9)/2.7(1.8)/2.8(1.6)	2.7(0.6)	5.3(9.2)/1.5(0.2)/1.7(0.3)	0(0)	3.8(1.4)/0.1(0)/0.1(0)	3.1(0.9)	5.4(10.4)/2(0.6)/2.1(0.7)	-
	p=180	1.9(0.5)	5.3(32.2)/1.8(10.9)/1.9(9.8)	1.1(0.1)	5.6(49)/0.5(0)/0.6(0)	0(0)	4.2(8.4)/0(0)/0(0)	1.4(0.2)	5.8(74.6)/0.7(0.1)/0.8(0.1)	-

Table S2: The performance of BS using different selection rules, Orth-Sparse-Ex1, n=2000

		C_p		AICc		BIC		GCV		CV
		edf	ndf/hdf/bdf	edf	ndf/hdf/bdf	edf	ndf/hdf/bdf	edf	ndf/hdf/bdf	
		% worse than the best possible BS								
hsnr	p=14	8	33/7/9	8	33/7/9	0	3/0/0	8	33/7/9	18
	p=30	3	85/3/6	3	85/3/6	0	9/0/0	3	86/3/6	23
	p=60	2	155/2/4	2	156/2/4	0	21/0/0	2	156/2/4	-
	p=180	0	334/1/3	1	337/1/3	0	60/0/0	1	340/1/3	-
msnr	p=14	8	33/7/9	8	33/7/9	0	3/0/0	8	33/7/9	18
	p=30	3	85/3/6	3	85/3/6	0	9/0/0	3	86/3/6	23
	p=60	2	155/2/4	2	156/2/4	0	21/0/0	2	156/2/4	-
	p=180	0	334/1/3	1	337/1/3	0	60/0/0	1	340/1/3	-
lsnr	p=14	8	33/9/9	8	33/9/9	0	3/0/0	8	33/9/9	18
	p=30	3	85/6/7	3	85/5/6	0	9/0/0	3	86/6/7	23
	p=60	2	155/5/5	2	156/5/5	0	21/0/0	2	156/5/5	-
	p=180	0	334/5/4	1	337/4/4	0	60/1/1	1	340/5/4	-
		Relative efficiency								
hsnr	p=14	0.93	0.75/0.94/0.92	0.93	0.75/0.94/0.92	1	0.97/1/1	0.92	0.75/0.94/0.92	0.84
	p=30	0.97	0.54/0.97/0.94	0.97	0.54/0.97/0.94	1	0.92/1/1	0.97	0.54/0.97/0.94	0.81
	p=60	0.98	0.39/0.98/0.96	0.98	0.39/0.98/0.96	1	0.83/1/1	0.98	0.39/0.98/0.96	-
	p=180	1	0.23/0.99/0.97	0.99	0.23/0.99/0.97	1	0.62/1/1	0.99	0.23/0.99/0.97	-
msnr	p=14	0.93	0.75/0.94/0.92	0.93	0.75/0.94/0.92	1	0.97/1/1	0.92	0.75/0.94/0.92	0.84
	p=30	0.97	0.54/0.97/0.94	0.97	0.54/0.97/0.94	1	0.92/1/1	0.97	0.54/0.97/0.94	0.81
	p=60	0.98	0.39/0.98/0.96	0.98	0.39/0.98/0.96	1	0.83/1/1	0.98	0.39/0.98/0.96	-
	p=180	1	0.23/0.99/0.97	0.99	0.23/0.99/0.97	1	0.62/1/1	0.99	0.23/0.99/0.97	-
lsnr	p=14	0.93	0.75/0.92/0.92	0.93	0.75/0.92/0.92	1	0.97/1/1	0.92	0.75/0.92/0.92	0.84
	p=30	0.97	0.54/0.95/0.94	0.97	0.54/0.95/0.94	1	0.92/1/1	0.97	0.54/0.95/0.94	0.81
	p=60	0.98	0.39/0.95/0.95	0.98	0.39/0.95/0.95	1	0.83/1/1	0.98	0.39/0.95/0.95	-
	p=180	1	0.23/0.96/0.96	0.99	0.23/0.96/0.96	1	0.62/0.99/0.99	0.99	0.23/0.96/0.96	-
		Sparsistency (number of extra variables)								
hsnr	p=14	6(0.3)	6(1.2)/6(0.3)/6(0.3)	6(0.3)	6(1.2)/6(0.3)/6(0.3)	6(0)	6(0)/6(0)/6(0)	6(0.3)	6(1.2)/6(0.3)/6(0.3)	6(0.6)
	p=30	6(0.1)	6(3.8)/6(0.1)/6(0.2)	6(0.1)	6(3.8)/6(0.1)/6(0.2)	6(0)	6(0.1)/6(0)/6(0)	6(0.1)	6(3.9)/6(0.1)/6(0.2)	6(0.6)
	p=60	6(0)	6(8.6)/6(0)/6(0)	6(0)	6(8.6)/6(0)/6(0)	6(0)	6(0.3)/6(0)/6(0)	6(0)	6(8.7)/6(0)/6(0)	-
	p=180	6(0)	6(27.5)/6(0)/6(0)	6(0)	6(28.2)/6(0)/6(0)	6(0)	6(1.1)/6(0)/6(0)	6(0)	6(28.9)/6(0)/6(0)	-
msnr	p=14	6(0.3)	6(1.2)/6(0.3)/6(0.3)	6(0.3)	6(1.2)/6(0.3)/6(0.3)	6(0)	6(0)/6(0)/6(0)	6(0.3)	6(1.2)/6(0.3)/6(0.3)	6(0.6)
	p=30	6(0.1)	6(3.8)/6(0.1)/6(0.2)	6(0.1)	6(3.8)/6(0.1)/6(0.2)	6(0)	6(0.1)/6(0)/6(0)	6(0.1)	6(3.9)/6(0.1)/6(0.2)	6(0.6)
	p=60	6(0)	6(8.6)/6(0)/6(0)	6(0)	6(8.6)/6(0)/6(0)	6(0)	6(0.3)/6(0)/6(0)	6(0)	6(8.7)/6(0)/6(0)	-
	p=180	6(0)	6(27.5)/6(0)/6(0)	6(0)	6(28.2)/6(0)/6(0)	6(0)	6(1.1)/6(0)/6(0)	6(0)	6(28.9)/6(0)/6(0)	-
lsnr	p=14	6(0.3)	6(1.2)/6(0.4)/6(0.3)	6(0.3)	6(1.2)/6(0.4)/6(0.3)	6(0)	6(0)/6(0)/6(0)	6(0.3)	6(1.2)/6(0.4)/6(0.3)	6(0.6)
	p=30	6(0.1)	6(3.8)/6(0.2)/6(0.2)	6(0.1)	6(3.8)/6(0.2)/6(0.2)	6(0)	6(0.1)/6(0)/6(0)	6(0.1)	6(3.9)/6(0.2)/6(0.2)	6(0.6)
	p=60	6(0)	6(8.6)/6(0.1)/6(0.1)	6(0)	6(8.6)/6(0.1)/6(0.1)	6(0)	6(0.3)/6(0)/6(0)	6(0)	6(8.7)/6(0.1)/6(0.1)	-
	p=180	6(0)	6(27.5)/6(0.1)/6(0)	6(0)	6(28.2)/6(0.1)/6(0)	6(0)	6(1.1)/6(0)/6(0)	6(0)	6(28.9)/6(0.1)/6(0)	-

Table S3: The performance of BS using different selection rules, Orth-Sparse-Ex2, n=200

		C_p		AICc		BIC		GCV		CV
		edf	ndf/hdf/bdf	edf	ndf/hdf/bdf	edf	ndf/hdf/bdf	edf	ndf/hdf/bdf	
		% worse than the best possible BS								
hsnr	p=14	23	21/32/29	23	21/32/28	39	20/40/38	24	21/32/29	23
	p=30	21	48/27/26	21	47/25/23	27	20/27/26	21	49/27/25	24
	p=60	17	89/20/19	17	91/18/18	19	33/19/18	17	96/18/18	-
	p=180	12	200/32/33	11	236/11/11	11	112/11/11	11	262/11/12	-
msnr	p=14	13	33/23/20	11	33/21/17	3	14/12/11	14	34/23/19	21
	p=30	6	78/21/19	4	78/15/14	1	28/19/18	6	80/18/17	24
	p=60	3	146/20/17	3	148/14/13	1	59/34/28	3	155/17/15	-
	p=180	3	314/50/50	2	365/15/15	2	184/65/57	3	400/16/16	-
lsnr	p=14	25	26/34/32	25	26/34/33	52	24/91/77	26	27/34/33	27
	p=30	21	54/37/34	20	54/34/31	48	23/90/79	22	56/35/33	29
	p=60	19	95/34/32	17	97/33/31	49	35/84/76	18	102/33/32	-
	p=180	15	198/57/54	15	235/39/35	56	105/72/68	15	260/37/34	-
		Relative efficiency								
hsnr	p=14	0.97	0.99/0.91/0.93	0.97	0.99/0.91/0.93	0.86	1/0.86/0.87	0.97	0.99/0.91/0.93	0.97
	p=30	0.99	0.81/0.94/0.95	0.99	0.81/0.96/0.97	0.94	1/0.94/0.95	0.99	0.8/0.95/0.96	0.97
	p=60	0.99	0.62/0.97/0.98	1	0.61/0.99/0.99	0.98	0.87/0.98/0.98	1	0.59/0.98/0.99	-
	p=180	0.99	0.37/0.84/0.84	1	0.33/1/1	1	0.52/1/1	1	0.31/1/0.99	-
msnr	p=14	0.91	0.77/0.83/0.85	0.92	0.77/0.85/0.87	1	0.9/0.92/0.93	0.9	0.77/0.83/0.86	0.85
	p=30	0.95	0.57/0.84/0.85	0.97	0.57/0.88/0.89	1	0.79/0.85/0.86	0.95	0.56/0.86/0.86	0.81
	p=60	0.98	0.41/0.84/0.86	0.99	0.41/0.88/0.89	1	0.64/0.76/0.79	0.98	0.4/0.86/0.88	-
	p=180	0.99	0.25/0.68/0.68	1	0.22/0.89/0.89	1	0.36/0.62/0.65	1	0.2/0.88/0.88	-
lsnr	p=14	0.99	0.98/0.92/0.93	0.99	0.98/0.92/0.93	0.81	1/0.65/0.7	0.98	0.97/0.92/0.93	0.97
	p=30	0.99	0.78/0.88/0.9	1	0.78/0.9/0.92	0.82	0.98/0.63/0.67	0.98	0.77/0.89/0.9	0.93
	p=60	0.99	0.6/0.88/0.89	1	0.6/0.88/0.9	0.79	0.87/0.64/0.67	0.99	0.58/0.88/0.89	-
	p=180	1	0.39/0.73/0.75	1	0.34/0.83/0.85	0.74	0.56/0.67/0.69	1	0.32/0.84/0.86	-
		Sparsistency (number of extra variables)								
hsnr	p=14	5.3(0.9)	5.6(1.3)/5.1(1.7)/5.2(1.2)	5.2(0.7)	5.6(1.2)/5(1.3)/5.1(1)	4.1(0)	5.1(0.2)/4.1(0)/4.2(0)	5.3(0.9)	5.6(1.3)/5.1(1.6)/5.2(1.2)	5.3(1)
	p=30	4.8(0.4)	5.6(3.9)/4.6(0.8)/4.7(0.7)	4.7(0.2)	5.6(3.8)/4.5(0.2)/4.6(0.2)	4(0)	5.1(0.6)/4(0)/4.1(0)	4.8(0.4)	5.7(4.1)/4.5(0.6)/4.6(0.6)	5(0.9)
	p=60	4.5(0.2)	5.6(8.9)/4.3(0.3)/4.4(0.3)	4.4(0.1)	5.7(9.2)/4.2(0.1)/4.3(0.1)	4(0)	5.1(1.5)/4(0)/4(0)	4.4(0.1)	5.7(10.5)/4.3(0.1)/4.4(0.1)	-
	p=180	4.2(0)	5.7(32.2)/4.3(7.8)/4.4(7.4)	4.1(0)	5.8(48.9)/4.1(0)/4.1(0)	4(0)	5.3(9.1)/4(0)/4(0)	4.2(0)	5.9(74.6)/4.1(0)/4.2(0.1)	-
msnr	p=14	4.2(0.4)	4.8(1.3)/4.5(1.2)/4.4(0.9)	4.2(0.3)	4.8(1.2)/4.4(1)/4.3(0.6)	4(0)	4.3(0.2)/4(0)/4(0)	4.2(0.5)	4.8(1.3)/4.5(1.2)/4.4(0.8)	4.4(0.7)
	p=30	4.1(0.1)	4.8(3.9)/4.2(0.7)/4.2(0.6)	4(0.1)	4.8(3.8)/4.1(0.3)/4.1(0.2)	4(0)	4.3(0.6)/3.8(0)/3.9(0)	4.1(0.2)	4.8(4.1)/4.2(0.5)/4.1(0.5)	4.2(0.7)
	p=60	4(0)	4.8(8.9)/4.1(0.3)/4.1(0.3)	4(0)	4.8(9.3)/4(0.2)/4(0.1)	4(0)	4.3(1.5)/3.7(0)/3.8(0)	4(0)	4.9(10.5)/4.1(0.2)/4(0.2)	-
	p=180	4(0)	4.8(32.2)/4.1(7.3)/4.1(7)	4(0)	5.1(49.2)/3.9(0.1)/3.9(0.1)	4(0)	4.4(8.9)/3.4(0)/3.5(0)	4(0)	5.3(74.7)/3.9(0.1)/3.9(0.1)	-
lsnr	p=14	3.4(1)	3.8(1.3)/3.9(2.4)/3.7(1.8)	3.3(0.8)	3.8(1.2)/3.7(2.1)/3.5(1.5)	1.8(0)	2.8(0.2)/1.1(0)/1.4(0)	3.4(1.1)	3.8(1.3)/3.9(2.4)/3.6(1.7)	3.2(0.8)
	p=30	2.7(0.6)	3.8(3.9)/2.8(2)/2.7(1.5)	2.6(0.4)	3.8(3.8)/2.5(0.9)/2.4(0.7)	1.5(0)	2.8(0.6)/0.6(0)/0.9(0)	2.7(0.7)	3.9(4.1)/2.7(1.4)/2.6(1.2)	2.8(1)
	p=60	2.3(0.3)	3.8(8.9)/2.2(0.9)/2.2(0.8)	2.2(0.2)	3.9(9.3)/1.8(0.2)/1.9(0.3)	1.2(0)	2.8(1.5)/0.4(0)/0.6(0)	2.2(0.3)	4(10.5)/2(0.4)/2(0.4)	-
	p=180	1.9(0.3)	3.9(32.2)/1.8(9)/1.9(8.2)	1.8(0.2)	4.3(49.6)/1.1(0.1)/1.2(0.1)	0.6(0)	3.1(8.4)/0.2(0)/0.3(0)	1.9(0.2)	4.7(74.7)/1.2(0.1)/1.3(0.1)	-

Table S4: The performance of BS using different selection rules, Orth-Sparse-Ex2, n=2000

		C_p		AICc		BIC		GCV		CV
		edf	ndf/hdf/bdf	edf	ndf/hdf/bdf	edf	ndf/hdf/bdf	edf	ndf/hdf/bdf	
		% worse than the best possible BS								
hsnr	p=14	8	33/10/9	8	33/9/9	0	3/0/1	8	33/9/9	18
	p=30	3	85/6/7	3	85/5/6	0	9/0/0	3	86/6/7	23
	p=60	2	155/4/5	2	156/4/5	0	21/1/1	2	156/4/5	-
	p=180	0	334/4/4	1	337/4/3	0	60/4/3	1	340/4/4	-
msnr	p=14	13	27/31/28	13	27/32/27	47	26/95/80	13	27/32/28	22
	p=30	14	66/39/35	13	66/39/35	74	23/93/86	13	66/39/35	31
	p=60	15	111/39/35	14	111/38/35	80	22/82/78	15	112/39/35	-
	p=180	15	217/37/35	16	219/37/35	63	34/63/62	16	221/37/35	-
lsnr	p=14	15	29/19/17	15	29/19/17	4	8/8/7	15	29/19/17	20
	p=30	7	71/13/12	7	71/13/12	1	11/10/7	7	71/13/12	24
	p=60	3	131/10/9	3	131/10/9	1	19/16/11	3	132/10/9	-
	p=180	1	288/9/8	1	291/8/7	0	51/35/30	1	293/9/8	-
		Relative efficiency								
hsnr	p=14	0.93	0.75/0.91/0.91	0.93	0.75/0.92/0.92	1	0.97/1/0.99	0.92	0.75/0.91/0.91	0.84
	p=30	0.97	0.54/0.95/0.94	0.97	0.54/0.95/0.94	1	0.92/1/1	0.97	0.54/0.95/0.94	0.81
	p=60	0.98	0.39/0.96/0.95	0.98	0.39/0.96/0.96	1	0.83/0.99/0.99	0.98	0.39/0.96/0.95	-
	p=180	1	0.23/0.96/0.97	0.99	0.23/0.97/0.97	1	0.62/0.96/0.97	0.99	0.23/0.96/0.97	-
msnr	p=14	1	0.89/0.86/0.88	1	0.89/0.86/0.89	0.77	0.9/0.58/0.63	1	0.89/0.86/0.88	0.93
	p=30	1	0.68/0.82/0.84	1	0.68/0.82/0.84	0.65	0.92/0.59/0.61	1	0.68/0.82/0.84	0.87
	p=60	1	0.54/0.83/0.85	1	0.54/0.83/0.85	0.64	0.94/0.63/0.64	1	0.54/0.83/0.85	-
	p=180	1	0.36/0.84/0.85	1	0.36/0.84/0.85	0.71	0.86/0.71/0.71	1	0.36/0.84/0.85	-
lsnr	p=14	0.9	0.81/0.88/0.89	0.91	0.81/0.88/0.89	1	0.96/0.97/0.97	0.91	0.81/0.88/0.89	0.87
	p=30	0.95	0.59/0.9/0.9	0.95	0.59/0.9/0.91	1	0.92/0.92/0.95	0.95	0.59/0.9/0.91	0.82
	p=60	0.98	0.44/0.91/0.92	0.98	0.43/0.92/0.93	1	0.84/0.87/0.9	0.98	0.43/0.91/0.93	-
	p=180	0.99	0.26/0.92/0.93	0.99	0.26/0.93/0.93	1	0.66/0.74/0.77	0.99	0.26/0.92/0.93	-
		Sparsistency (number of extra variables)								
hsnr	p=14	6(0.3)	6(1.2)/6(0.4)/6(0.3)	6(0.3)	6(1.2)/6(0.4)/6(0.3)	6(0)	6(0)/6(0)/6(0)	6(0.3)	6(1.2)/6(0.4)/6(0.3)	6(0.6)
	p=30	6(0.1)	6(3.8)/6(0.2)/6(0.2)	6(0.1)	6(3.8)/6(0.2)/6(0.2)	6(0)	6(0.1)/6(0)/6(0)	6(0.1)	6(3.9)/6(0.2)/6(0.2)	6(0.6)
	p=60	6(0)	6(8.6)/6(0.1)/6(0.1)	6(0)	6(8.6)/6(0.1)/6(0.1)	6(0)	6(0.3)/6(0)/6(0)	6(0)	6(8.7)/6(0.1)/6(0.1)	-
	p=180	6(0)	6(27.5)/6(0)/6(0)	6(0)	6(28.2)/6(0)/6(0)	6(0)	6(1.1)/6(0)/6(0)	6(0)	6(28.9)/6(0)/6(0)	-
msnr	p=14	5.9(0.6)	6(1.2)/5.8(1.7)/5.8(1.2)	5.9(0.6)	6(1.2)/5.8(1.7)/5.8(1.1)	5.1(0)	5.5(0)/4.4(0)/4.6(0)	5.9(0.6)	6(1.2)/5.8(1.7)/5.8(1.2)	5.9(0.8)
	p=30	5.8(0.4)	6(3.8)/5.5(1)/5.6(0.8)	5.8(0.4)	6(3.8)/5.5(0.9)/5.5(0.8)	4.5(0)	5.5(0.1)/4.2(0)/4.3(0)	5.8(0.4)	6(3.9)/5.5(1)/5.6(0.8)	5.7(1)
	p=60	5.6(0.4)	6(8.6)/5.2(0.3)/5.3(0.3)	5.6(0.4)	6(8.6)/5.2(0.3)/5.3(0.3)	4.1(0)	5.5(0.3)/4.1(0)/4.2(0)	5.6(0.4)	6(8.7)/5.2(0.3)/5.3(0.3)	-
	p=180	5.4(0.3)	6(27.5)/4.8(0.1)/4.8(0.1)	5.4(0.3)	6(28.2)/4.8(0.1)/4.8(0.1)	4(0)	5.5(1.1)/4(0)/4.1(0)	5.4(0.3)	6(28.9)/4.8(0.1)/4.8(0.1)	-
lsnr	p=14	4.3(0.5)	4.9(1.2)/4.5(1)/4.4(0.6)	4.3(0.5)	4.9(1.2)/4.5(0.9)/4.4(0.6)	4(0)	4.1(0)/4(0)/4(0)	4.3(0.5)	4.9(1.2)/4.5(1)/4.4(0.6)	4.5(0.7)
	p=30	4.1(0.2)	4.9(3.8)/4.2(0.4)/4.1(0.4)	4.1(0.2)	4.9(3.8)/4.2(0.4)/4.1(0.3)	4(0)	4.1(0.1)/3.9(0)/4(0)	4.1(0.2)	4.9(3.9)/4.2(0.4)/4.1(0.3)	4.3(0.8)
	p=60	4(0)	4.9(8.6)/4.1(0.2)/4.1(0.1)	4(0)	4.9(8.6)/4.1(0.1)/4.1(0.1)	4(0)	4.1(0.3)/3.9(0)/3.9(0)	4(0)	4.9(8.7)/4.1(0.2)/4.1(0.1)	-
	p=180	4(0)	4.9(27.5)/4(0.1)/4(0.1)	4(0)	4.9(28.2)/4(0.1)/4(0.1)	4(0)	4.1(1.1)/3.7(0)/3.8(0)	4(0)	4.9(28.9)/4(0.1)/4(0.1)	-

Table S5: The performance of BS using different selection rules, Orth-Dense, n=200

		C_p		AICc		BIC		GCV		CV
		edf	ndf/hdf/bdf	edf	ndf/hdf/bdf	edf	ndf/hdf/bdf	edf	ndf/hdf/bdf	
		% worse than the best possible BS								
hsnr	p=14	0	0/0/0	0	0/0/0	0	1/0/0	0	0/0/0	0
	p=30	1	11/1/2	1	13/1/2	1	28/3/5	1	11/1/2	7
	p=60	8	7/9/9	9	7/11/11	20	8/32/33	8	8/10/10	-
	p=180	7	45/21/20	9	52/18/19	18	26/39/42	7	64/13/13	-
msnr	p=14	0	9/0/1	0	10/0/1	0	36/1/2	0	9/0/1	6
	p=30	3	10/3/4	3	11/4/5	21	27/19/25	3	10/4/4	11
	p=60	10	11/14/13	10	11/13/13	26	10/48/48	10	12/14/13	-
	p=180	8	52/23/23	10	62/18/19	21	25/61/56	8	74/14/14	-
lsnr	p=14	5	22/6/8	7	23/8/10	73	50/73/72	6	22/7/8	19
	p=30	15	10/16/16	20	10/21/20	27	16/27/27	17	10/18/18	16
	p=60	13	25/17/16	13	25/13/13	13	11/13/13	13	26/14/14	-
	p=180	8	86/22/22	7	102/7/7	7	39/7/7	7	116/7/7	-
		Relative efficiency								
hsnr	p=14	1	1/1/1	1	1/1/1	1	0.99/1/1	1	1/1/1	1
	p=30	1	0.91/1/1	1	0.9/1/0.99	1	0.79/0.98/0.96	1	0.91/1/1	0.95
	p=60	0.99	1/0.98/0.98	0.98	1/0.97/0.96	0.89	0.99/0.81/0.8	0.99	0.99/0.98/0.98	-
	p=180	1	0.74/0.89/0.89	0.99	0.71/0.91/0.9	0.91	0.85/0.77/0.76	1	0.65/0.95/0.95	-
msnr	p=14	1	0.92/1/0.99	1	0.91/1/0.99	1	0.74/1/0.99	1	0.92/1/0.99	0.95
	p=30	1	0.93/0.99/0.99	0.99	0.92/0.98/0.98	0.85	0.81/0.87/0.82	1	0.93/0.99/0.99	0.93
	p=60	1	0.99/0.96/0.97	1	0.99/0.97/0.97	0.87	1/0.74/0.74	1	0.98/0.97/0.97	-
	p=180	1	0.71/0.88/0.88	0.98	0.67/0.91/0.91	0.89	0.87/0.67/0.69	1	0.62/0.95/0.95	-
lsnr	p=14	0.98	0.85/0.97/0.96	0.97	0.84/0.96/0.94	0.6	0.69/0.6/0.6	0.98	0.85/0.97/0.95	0.86
	p=30	0.95	1/0.95/0.95	0.91	1/0.91/0.91	0.86	0.94/0.86/0.86	0.93	1/0.93/0.93	0.94
	p=60	0.98	0.89/0.95/0.96	0.99	0.89/0.99/0.99	0.98	1/0.98/0.98	0.98	0.88/0.97/0.98	-
	p=180	1	0.58/0.88/0.88	1	0.53/1/1	1	0.77/1/1	1	0.5/1/1	-
		Sparsistency (number of extra variables)								
hsnr	p=14	14	14/14/14	14	14/14/14	14	14/14/14	14	14/14/14	14
	p=30	30	24.7/29.5/29	30	24.2/29.4/28.8	30	20.9/28.8/27.5	30	24.7/29.5/29	26.6
	p=60	29.8	30.5/38.4/35.8	22.2	29.4/25.6/24.5	17.8	22.5/16.8/16.5	28.6	31.3/36.8/34	-
	p=180	20.5	53.3/37.4/35.5	18.3	62.3/16.3/16.3	16.1	35/13.7/13.5	19.4	89.8/17.8/17.8	-
msnr	p=14	14	13.2/14/13.9	14	13.2/14/13.9	14	11.8/13.9/13.8	14	13.2/14/13.9	13.4
	p=30	27.3	18.8/27.4/26.1	26.5	18.3/26.8/25.3	18	13.4/20.4/17.6	27.3	18.8/27.4/26.1	20.8
	p=60	19.4	24.1/29.6/27	13.9	23.4/15.6/15.2	9.3	14.5/7.5/7.4	18.3	25.2/26/24.1	-
	p=180	12.6	47.1/29.1/28.1	10.4	59/8.8/8.8	8.1	24.4/4.8/5	11.3	86.4/10/10	-
lsnr	p=14	13.6	7.7/12.7/11.7	13.4	7.6/12.3/11.3	0.7	3.6/0.7/0.8	13.5	7.8/12.6/11.6	8.8
	p=30	12.8	10.5/14.6/13	7.6	10.3/8.5/7.6	0	4/0/0	11.3	10.8/12.3/11.2	7.5
	p=60	3.4	15.7/6.5/6	1	15.8/0.8/1	0	4.9/0/0	2	17.3/2.4/2.4	-
	p=180	0.8	39/14.5/13.7	0.3	55.2/0.2/0.3	0	11.8/0/0	0.4	81.7/0.3/0.4	-

Table S6: The performance of BS using different selection rules, Orth-Dense, n=2000

		C_p		AICc		BIC		GCV		CV
		edf	ndf/hdf/bdf	edf	ndf/hdf/bdf	edf	ndf/hdf/bdf	edf	ndf/hdf/bdf	
		% worse than the best possible BS								
hsnr	p=14	0	0/0/0	0	0/0/0	0	0/0/0	0	0/0/0	0
	p=30	0	1/0/0	0	1/0/0	0	18/0/1	0	1/0/0	1
	p=60	5	5/5/5	5	5/5/5	25	17/34/37	5	5/5/5	-
	p=180	6	34/8/8	6	34/8/8	19	7/36/37	6	35/8/8	-
msnr	p=14	0	0/0/0	0	0/0/0	0	0/0/0	0	0/0/0	0
	p=30	1	9/1/1	1	9/1/1	1	40/2/5	1	9/1/1	5
	p=60	7	6/8/8	7	6/8/8	28	15/37/40	7	6/8/8	-
	p=180	6	39/9/9	6	39/8/9	21	7/38/40	6	40/8/9	-
lsnr	p=14	0	5/0/0	0	5/0/0	0	49/0/1	0	5/0/0	4
	p=30	2	11/3/3	2	11/3/3	44	41/36/45	2	11/3/3	10
	p=60	10	10/13/12	10	10/13/12	32	16/45/48	10	10/13/12	-
	p=180	8	48/10/10	8	48/10/10	24	8/45/47	8	49/10/10	-
		Relative efficiency								
hsnr	p=14	1	1/1/1	1	1/1/1	1	1/1/1	1	1/1/1	1
	p=30	1	0.99/1/1	1	0.99/1/1	1	0.85/1/0.99	1	0.99/1/1	0.99
	p=60	0.99	1/0.99/0.99	0.99	1/0.99/0.99	0.83	0.89/0.78/0.76	0.99	1/0.99/0.99	-
	p=180	1	0.79/0.98/0.98	1	0.79/0.98/0.98	0.89	1/0.78/0.78	1	0.79/0.98/0.98	-
msnr	p=14	1	1/1/1	1	1/1/1	1	1/1/1	1	1/1/1	1
	p=30	1	0.92/1/1	1	0.92/1/1	1	0.72/0.99/0.96	1	0.92/1/1	0.96
	p=60	0.99	1/0.98/0.98	0.99	1/0.98/0.98	0.83	0.92/0.77/0.76	0.99	1/0.98/0.98	-
	p=180	1	0.76/0.98/0.98	1	0.76/0.98/0.98	0.88	1/0.77/0.76	1	0.76/0.98/0.98	-
lsnr	p=14	1	0.95/1/1	1	0.95/1/1	1	0.67/1/0.99	1	0.95/1/1	0.96
	p=30	1	0.92/0.99/0.99	1	0.92/0.99/0.99	0.71	0.73/0.75/0.7	1	0.92/0.99/0.99	0.93
	p=60	1	1/0.97/0.98	1	1/0.97/0.98	0.83	0.94/0.75/0.74	1	1/0.97/0.98	-
	p=180	1	0.73/0.98/0.98	1	0.73/0.98/0.98	0.87	1/0.74/0.73	1	0.72/0.98/0.98	-
		Sparsistency (number of extra variables)								
hsnr	p=14	14	14/14/14	14	14/14/14	14	14/14/14	14	14/14/14	14
	p=30	30	29.8/30/29.9	30	29.8/30/29.9	30	28.6/30/29.9	30	29.8/30/29.9	29.8
	p=60	44.9	39.8/50.9/48.7	44.2	39.7/50.4/48.2	28.5	30.5/27.6/27.4	45.1	39.8/50.8/48.6	-
	p=180	32.1	58.9/32.4/32.3	31.8	58.9/31.6/31.6	27	31.3/25/24.9	32	59.9/32.1/32	-
msnr	p=14	14	14/14/14	14	14/14/14	14	14/14/14	14	14/14/14	14
	p=30	30	27.1/29.9/29.6	30	27.1/29.9/29.6	30	22.5/29.5/28.6	30	27.1/29.9/29.6	28.2
	p=60	34.8	33.3/42.8/40.1	33.9	33.2/41.9/39.2	20.4	22.9/19.4/19.2	34.6	33.3/42.8/40	-
	p=180	24.2	52.4/24.4/24.3	24	52.6/23.6/23.6	19.2	23.6/17.3/17.2	24.1	53.5/24.1/24.1	-
lsnr	p=14	14	13.6/14/13.9	14	13.6/14/13.9	14	11.7/14/13.9	14	13.6/14/13.9	13.7
	p=30	28.8	19.9/28.2/26.9	28.8	19.9/28.1/26.8	13.5	12.5/16.7/14.1	28.8	19.9/28.2/26.9	22.3
	p=60	21.6	24.8/30.6/27.9	20.9	24.7/29.2/26.9	10	12.7/8.7/8.5	21.8	24.9/30.4/27.7	-
	p=180	13.9	43.8/14/14	13.6	44.1/13.3/13.3	9.1	13.4/7/6.8	13.8	45/13.7/13.6	-

Table S7: The performance of BS compared to regularization methods, Orth-Sparse-Ex1, n=200

		BS AICc-hdf	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BS				
hsnr	p=14	6	42/41	16/19	13	18
	p=30	2	70/68	28/22	14	24
	p=60	2	94/92	53/25	17	28
	p=180	1	128/132	129/28	20	27
msnr	p=14	11	43/41	21/22	13	18
	p=30	8	70/68	44/26	16	24
	p=60	8	95/92	82/30	18	28
	p=180	7	127/132	228/32	19	28
lsnr	p=14	24	11/11	15/17	20	19
	p=30	37	7/7	16/13	15	15
	p=60	31	3/3	27/9	9	10
	p=180	22	1/5	99/8	8	10
		Relative efficiency				
hsnr	p=14	1	0.74/0.75	0.92/0.89	0.94	0.9
	p=30	1	0.6/0.61	0.8/0.84	0.9	0.83
	p=60	1	0.52/0.53	0.66/0.81	0.87	0.8
	p=180	1	0.44/0.43	0.44/0.79	0.84	0.79
msnr	p=14	1	0.78/0.79	0.92/0.91	0.98	0.94
	p=30	1	0.64/0.64	0.75/0.86	0.94	0.87
	p=60	1	0.56/0.56	0.6/0.83	0.92	0.85
	p=180	1	0.47/0.46	0.33/0.81	0.9	0.84
lsnr	p=14	0.89	1/1	0.97/0.95	0.92	0.93
	p=30	0.78	1/1	0.92/0.95	0.93	0.93
	p=60	0.78	1/1	0.81/0.95	0.94	0.93
	p=180	0.83	1/0.96	0.51/0.93	0.94	0.92
		Sparsistency (number of extra variables)				
hsnr	p=14	6(0.2)	6(3.8)/6(4.5)	6(0.9)/6(1.3)	6(0.6)	6(0.7)
	p=30	6(0.1)	6(7.7)/6(8.6)	6(2.1)/6(1.6)	6(1.1)	6(0.9)
	p=60	6(0)	6(11.1)/6(12.7)	6(5.3)/6(2)	6(1.7)	6(1)
	p=180	6(0)	6(14.3)/6(22.9)	6(20.5)/6(3.3)	6(3.7)	6(1.2)
msnr	p=14	6(0.5)	6(3.8)/6(4.5)	6(1.1)/6(1.3)	6(0.5)	6(0.7)
	p=30	6(0.2)	6(7.7)/6(8.6)	6(2.8)/6(1.4)	6(0.8)	6(0.9)
	p=60	6(0.1)	6(11.1)/6(12.6)	6(6.5)/6(1.8)	6(1.2)	6(1)
	p=180	6(0.1)	6(14.3)/6(23)	6(33.6)/6(2.7)	6(2.4)	6(1.2)
lsnr	p=14	5.4(4.2)	5.7(3.5)/5.7(4.1)	5.2(1.4)/5.3(2.9)	5.3(2.7)	5(1.8)
	p=30	3.3(2.2)	5.3(6.7)/5.4(7.1)	5.2(3.9)/4.9(4.8)	4.9(5.1)	4.5(3.3)
	p=60	1.5(0.2)	4.9(9.2)/4.9(9.9)	5.1(8.6)/4.4(6.7)	4.4(7.4)	4(4.7)
	p=180	0.5(0)	3.9(9.6)/4(14.8)	5.5(45.9)/3.5(10.5)	3.6(11.3)	3.1(6.8)

Table S8: The performance of BS compared to regularization methods, Orth-Sparse-Ex1, n=2000

		BS AICc-hdf	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BS				
hsnr	p=14	7	40/40	13/15	13	18
	p=30	3	70/68	21/17	14	24
	p=60	2	94/91	29/18	14	27
	p=180	1	130/125	52/19	14	25
msnr	p=14	7	41/40	15/17	13	18
	p=30	3	70/68	28/19	14	24
	p=60	2	94/90	44/21	15	27
	p=180	1	130/126	95/22	14	25
lsnr	p=14	9	41/40	21/21	14	18
	p=30	5	70/68	45/25	15	24
	p=60	5	94/90	81/27	15	27
	p=180	4	129/126	188/26	14	25
		Relative efficiency				
hsnr	p=14	1	0.76/0.76	0.95/0.92	0.95	0.9
	p=30	1	0.61/0.62	0.86/0.88	0.91	0.83
	p=60	1	0.53/0.54	0.79/0.86	0.9	0.81
	p=180	1	0.44/0.45	0.67/0.85	0.89	0.81
msnr	p=14	1	0.76/0.76	0.92/0.91	0.94	0.9
	p=30	1	0.61/0.62	0.81/0.87	0.91	0.83
	p=60	1	0.53/0.54	0.71/0.84	0.89	0.81
	p=180	1	0.44/0.45	0.52/0.84	0.89	0.81
lsnr	p=14	1	0.77/0.78	0.9/0.9	0.96	0.92
	p=30	1	0.62/0.63	0.73/0.85	0.92	0.85
	p=60	1	0.54/0.55	0.58/0.83	0.91	0.83
	p=180	1	0.46/0.46	0.36/0.83	0.91	0.84
		Sparsistency (number of extra variables)				
hsnr	p=14	6(0.3)	6(3.9)/6(4.4)	6(0.9)/6(1.3)	6(0.6)	6(0.7)
	p=30	6(0.1)	6(8.5)/6(8.7)	6(2.2)/6(1.8)	6(1.1)	6(0.8)
	p=60	6(0)	6(13.2)/6(12.4)	6(4)/6(2.1)	6(1.5)	6(0.8)
	p=180	6(0)	6(21.8)/6(18.7)	6(10.6)/6(2.7)	6(2.2)	6(0.8)
msnr	p=14	6(0.3)	6(3.9)/6(4.4)	6(1)/6(1.4)	6(0.6)	6(0.7)
	p=30	6(0.1)	6(8.6)/6(8.7)	6(2.5)/6(1.7)	6(1.1)	6(0.8)
	p=60	6(0)	6(13.3)/6(12.4)	6(5)/6(1.8)	6(1.4)	6(0.8)
	p=180	6(0)	6(21.8)/6(18.7)	6(15.6)/6(1.9)	6(2)	6(0.8)
lsnr	p=14	6(0.4)	6(3.9)/6(4.4)	6(1.1)/6(1.4)	6(0.6)	6(0.7)
	p=30	6(0.2)	6(8.5)/6(8.7)	6(3)/6(1.5)	6(0.8)	6(0.8)
	p=60	6(0.1)	6(13.2)/6(12.4)	6(6.7)/6(1.4)	6(0.9)	6(0.8)
	p=180	6(0.1)	6(21.2)/6(18.7)	6(23.6)/6(1.3)	6(1.3)	6(0.8)

Table S9: The performance of BS compared to regularization methods, Orth-Sparse-Ex2, n=200

		BS AICc-hdf	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BS				
hsnr	p=14	32	23/22	16/19	22	22
	p=30	25	32/30	32/20	19	23
	p=60	18	39/38	59/20	16	24
	p=180	11	51/55	189/19	14	23
msnr	p=14	21	32/31	25/19	17	20
	p=30	15	52/51	43/23	18	26
	p=60	14	70/69	73/27	22	29
	p=180	15	93/99	293/28	26	32
lsnr	p=14	34	19/18	18/24	24	23
	p=30	34	21/21	30/26	25	24
	p=60	33	23/24	57/26	25	25
	p=180	39	24/30	167/26	25	27
		Relative efficiency				
hsnr	p=14	0.88	0.95/0.95	1/0.98	0.95	0.95
	p=30	0.95	0.9/0.91	0.9/0.99	1	0.96
	p=60	0.99	0.84/0.84	0.73/0.97	1	0.94
	p=180	1	0.74/0.72	0.38/0.94	0.97	0.9
msnr	p=14	0.97	0.89/0.89	0.93/0.98	1	0.97
	p=30	1	0.76/0.76	0.81/0.94	0.98	0.92
	p=60	1	0.67/0.67	0.66/0.9	0.94	0.89
	p=180	1	0.59/0.58	0.29/0.9	0.91	0.87
lsnr	p=14	0.88	1/1	1/0.96	0.95	0.96
	p=30	0.9	1/1	0.93/0.96	0.97	0.98
	p=60	0.92	1/1	0.78/0.98	0.99	0.99
	p=180	0.9	1/0.96	0.47/0.99	0.99	0.98
		Sparsistency (number of extra variables)				
hsnr	p=14	5(1.3)	5.9(3.7)/5.9(4.4)	5.6(1.3)/5.4(1.5)	5.3(1.2)	5.3(1.1)
	p=30	4.5(0.2)	5.7(7.2)/5.8(8.2)	5.5(3.4)/5.1(2)	5.1(2)	5.1(1.5)
	p=60	4.2(0.1)	5.6(10.5)/5.7(11.8)	5.5(7.8)/4.9(2.6)	5(3)	4.8(1.8)
	p=180	4.1(0)	5.4(12.9)/5.4(20.3)	5.8(45.9)/4.6(3.7)	4.7(5.3)	4.5(2.3)
msnr	p=14	4.4(1)	5.2(3.2)/5.3(3.8)	4.7(1.1)/4.5(0.9)	4.4(0.8)	4.5(0.8)
	p=30	4.1(0.3)	5(6.4)/5(7.1)	4.6(2.9)/4.3(1.2)	4.3(1.1)	4.3(1.1)
	p=60	4(0.2)	4.8(9.4)/4.8(10.4)	4.6(6.7)/4.2(1.4)	4.2(1.9)	4.2(1.3)
	p=180	3.9(0.1)	4.5(12)/4.5(18.2)	5(46.3)/4.1(2.1)	4.1(3.7)	4.1(1.8)
lsnr	p=14	3.7(2.1)	4.3(2.7)/4.4(3)	3.7(1.3)/3.5(1.4)	3.5(1.4)	3.4(1.1)
	p=30	2.5(0.9)	3.9(5.1)/3.9(5.4)	3.7(3.4)/3(2.3)	3.1(2.7)	3(1.9)
	p=60	1.8(0.2)	3.6(7.7)/3.6(8)	3.7(8)/2.8(3.6)	3(4.7)	2.8(2.9)
	p=180	1.1(0.1)	3(9.6)/3.1(13.7)	4.1(43.1)/2.4(6.1)	2.6(7.7)	2.4(4.6)

Table S10: The performance of BS compared to regularization methods, Orth-Sparse-Ex2, n=2000

		BS AICc-hdf	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BS				
hsnr	p=14	9	40/40	20/16	14	19
	p=30	5	71/68	45/21	15	24
	p=60	4	95/91	83/24	14	27
	p=180	4	129/125	195/23	14	25
msnr	p=14	32	34/33	20/21	23	21
	p=30	39	52/50	44/28	29	29
	p=60	38	60/57	74/28	29	32
	p=180	37	67/64	155/26	25	31
lsnr	p=14	19	27/26	24/18	16	19
	p=30	13	46/44	57/20	15	22
	p=60	10	63/60	102/21	15	24
	p=180	8	88/84	230/19	15	23
		Relative efficiency				
hsnr	p=14	1	0.78/0.78	0.91/0.94	0.96	0.92
	p=30	1	0.62/0.63	0.73/0.87	0.91	0.85
	p=60	1	0.54/0.55	0.57/0.84	0.91	0.82
	p=180	1	0.45/0.46	0.35/0.84	0.91	0.83
msnr	p=14	0.91	0.9/0.9	1/0.99	0.98	0.99
	p=30	0.92	0.85/0.86	0.89/1	1	0.99
	p=60	0.93	0.8/0.82	0.74/1	1	0.97
	p=180	0.91	0.75/0.76	0.49/0.99	1	0.95
lsnr	p=14	0.97	0.91/0.92	0.94/0.98	1	0.97
	p=30	1	0.77/0.78	0.72/0.94	0.98	0.92
	p=60	1	0.67/0.69	0.54/0.91	0.95	0.88
	p=180	1	0.58/0.59	0.33/0.91	0.95	0.88
		Sparsistency (number of extra variables)				
hsnr	p=14	6(0.4)	6(3.8)/6(4.4)	6(1.1)/6(0.9)	6(0.6)	6(0.7)
	p=30	6(0.2)	6(8.5)/6(8.7)	6(3)/6(1.3)	6(0.9)	6(0.8)
	p=60	6(0.1)	6(13.2)/6(12.4)	6(6.8)/6(1.5)	6(1)	6(0.9)
	p=180	6(0)	6(21.5)/6(18.7)	6(24)/6(1.6)	6(1.4)	6(0.9)
msnr	p=14	5.8(1.7)	6(3.8)/6(4.4)	6(1.2)/5.9(1.5)	5.9(1.1)	5.9(0.9)
	p=30	5.5(0.9)	6(8.5)/6(8.6)	6(3.6)/5.8(2.3)	5.8(2.2)	5.8(1.3)
	p=60	5.2(0.3)	6(13.1)/6(12.2)	5.9(7.9)/5.7(2.5)	5.7(2.8)	5.6(1.4)
	p=180	4.8(0.1)	5.9(21.2)/5.9(18.3)	6(27.1)/5.5(3.2)	5.5(4.3)	5.3(1.9)
lsnr	p=14	4.5(0.9)	5.4(3.4)/5.4(3.9)	4.9(1.3)/4.6(1.1)	4.5(0.7)	4.6(0.8)
	p=30	4.2(0.4)	5.1(7.3)/5.2(7.3)	4.8(3.8)/4.4(1.2)	4.3(1.2)	4.4(1.1)
	p=60	4.1(0.1)	5(11.1)/4.9(10)	4.8(8.1)/4.2(1.2)	4.2(1.3)	4.2(1.1)
	p=180	4(0.1)	4.7(18.1)/4.6(15)	4.9(27.3)/4.1(1)	4.1(1.5)	4.1(1.2)

Table S11: The performance of BS compared to regularization methods, Orth-Dense, n=200

		BS AICc-hdf	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BS				
hsnr	p=14	0	0/0	0/1	0	64
	p=30	1	1/1	7/2	4	6
	p=60	11	2/0	-1/0	2	5
	p=180	18	23/15	4/7	7	13
msnr	p=14	0	1/1	6/1	3	21
	p=30	4	1/0	8/3	5	6
	p=60	13	0/-1	0/1	3	5
	p=180	18	13/10	18/7	7	10
lsnr	p=14	8	5/4	17/8	11	14
	p=30	21	-6/-6	4/0	-2	0
	p=60	13	-6/-6	13/-1	-3	0
	p=180	7	-2/0	73/3	2	4
		Relative efficiency				
hsnr	p=14	1	1/1	1/0.99	1	0.61
	p=30	1	1/1	0.94/1	0.97	0.95
	p=60	0.9	0.97/0.99	1/0.99	0.97	0.94
	p=180	0.88	0.84/0.9	1/0.97	0.97	0.92
msnr	p=14	1	1/1	0.95/0.99	0.97	0.83
	p=30	0.96	0.99/1	0.93/0.97	0.95	0.94
	p=60	0.87	0.98/1	0.99/0.97	0.96	0.94
	p=180	0.9	0.95/0.97	0.91/1	1	0.97
lsnr	p=14	0.96	0.99/0.99	0.88/0.96	0.93	0.91
	p=30	0.78	1/1	0.9/0.94	0.95	0.94
	p=60	0.83	1/1	0.83/0.95	0.96	0.94
	p=180	0.91	1/0.98	0.56/0.95	0.96	0.94
		Sparsistency (number of extra variables)				
hsnr	p=14	14	14/14	14/14	14	13
	p=30	29.4	28.9/29.4	26.5/29.1	27.8	26.5
	p=60	25.6	41.5/46.2	30.7/38.9	35.1	30.1
	p=180	16.3	41.1/68.3	39.9/34.1	32.7	22.9
msnr	p=14	14	13.9/14	13.5/13.9	13.6	12.6
	p=30	26.8	25.5/26.9	19.1/25	23.9	21.5
	p=60	15.6	32.8/37.5	23.5/29	28.3	22.5
	p=180	8.8	30.7/48.8	44/26.6	25.3	17.2
lsnr	p=14	12.3	10.9/11.5	8.2/11	10.4	9
	p=30	8.5	14.1/14.8	10.9/13.2	13.3	11
	p=60	0.8	14.6/16	16.1/12.8	14.1	10.6
	p=180	0.2	10.5/15.3	51.1/12.7	13.6	8.7

Table S12: The performance of BS compared to regularization methods, Orth-Dense, n=2000

		BS AICc-hdf	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BS				
hsnr	p=14	0	2/2	1/4	0	303
	p=30	0	1/1	3/6	1	16
	p=60	5	1/1	0/0	2	5
	p=180	8	17/16	4/5	3	9
msnr	p=14	0	0/0	0/2	0	113
	p=30	1	1/1	3/2	3	7
	p=60	8	2/1	0/0	3	5
	p=180	8	15/14	4/4	4	9
lsnr	p=14	0	0/0	3/1	3	27
	p=30	3	1/0	7/2	5	6
	p=60	13	0/-1	0/1	3	4
	p=180	10	9/8	14/5	5	8
		Relative efficiency				
hsnr	p=14	1	0.98/0.98	0.99/0.96	1	0.25
	p=30	1	0.99/0.99	0.97/0.94	0.99	0.86
	p=60	0.95	0.99/0.99	1/1	0.97	0.95
	p=180	0.96	0.88/0.89	0.99/0.98	1	0.95
msnr	p=14	1	1/1	1/0.99	1	0.47
	p=30	1	1/1	0.98/0.99	0.98	0.94
	p=60	0.92	0.98/0.99	1/0.99	0.97	0.95
	p=180	0.96	0.9/0.91	1/0.99	1	0.95
lsnr	p=14	1	1/1	0.98/0.99	0.98	0.79
	p=30	0.98	1/1	0.94/0.98	0.96	0.94
	p=60	0.88	0.99/1	0.99/0.99	0.97	0.95
	p=180	0.95	0.96/0.97	0.92/0.99	1	0.97
		Sparsistency (number of extra variables)				
hsnr	p=14	14	14/14	14/14	14	13
	p=30	30	30/30	30/30	29.9	28.8
	p=60	50.4	53.5/54.6	46.8/50.4	45.4	41.4
	p=180	31.6	89.1/89.8	62.4/60.1	46.3	35.3
msnr	p=14	14	14/14	14/14	14	13
	p=30	29.9	29.6/29.8	29/29.6	28.8	27.7
	p=60	41.9	48.7/49.7	37.2/44.3	38.2	33.7
	p=180	23.6	75.7/75.6	51.1/43.5	37.6	28.2
lsnr	p=14	14	14/14	13.8/14	13.8	12.7
	p=30	28.1	26.8/27.6	21.2/26.3	24.8	22.7
	p=60	29.2	38.4/39.1	25.8/31	29.5	24
	p=180	13.3	55.1/53.1	42.6/30.7	29.3	21.7

Table S13: The performance of BOSS compared to other methods, Sparse-Ex1, ρ=0, n=200

		BOSS C_p-hdf/AICc-hdf/CV	BS CV	FS CV	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BOSS						
hsnr	p=14	8/6/18	19	19	42/41	16/20	13	18
	p=30	5/3/23	25	23	71/69	32/23	14	24
	p=60	4/2/21	-	23	87/85	51/24	16	25
	p=180	34/1/19	-	22	119/121	134/25	17	25
msnr	p=14	17/14/18	19	19	43/42	23/23	14	18
	p=30	15/11/23	25	23	71/69	49/28	16	24
	p=60	13/9/22	-	24	87/85	82/28	17	26
	p=180	44/7/20	-	22	119/121	222/30	17	27
lsnr	p=14	22/24/25	26	25	8/9	13/15	18	17
	p=30	32/34/26	26	26	2/2	14/9	10	9
	p=60	27/29/24	-	24	-1/-2	27/5	6	5
	p=180	32/22/18	-	19	-4/-2	83/2	1	2
		Relative efficiency						
hsnr	p=14	0.98/1/0.89	0.89	0.89	0.74/0.75	0.91/0.88	0.93	0.9
	p=30	0.98/1/0.84	0.82	0.83	0.6/0.61	0.78/0.83	0.9	0.83
	p=60	0.99/1/0.84	-	0.83	0.55/0.55	0.68/0.83	0.88	0.82
	p=180	0.75/1/0.85	-	0.83	0.46/0.46	0.43/0.81	0.87	0.81
msnr	p=14	0.98/1/0.96	0.96	0.96	0.8/0.81	0.93/0.93	1	0.96
	p=30	0.96/1/0.9	0.89	0.9	0.65/0.66	0.75/0.87	0.96	0.9
	p=60	0.97/1/0.9	-	0.88	0.58/0.59	0.6/0.85	0.94	0.87
	p=180	0.75/1/0.9	-	0.88	0.49/0.49	0.33/0.83	0.91	0.85
lsnr	p=14	0.89/0.88/0.87	0.86	0.86	1/1	0.95/0.94	0.92	0.93
	p=30	0.78/0.76/0.81	0.81	0.81	1/1	0.9/0.94	0.93	0.94
	p=60	0.77/0.76/0.79	-	0.79	1/1	0.78/0.94	0.93	0.94
	p=180	0.73/0.79/0.81	-	0.81	1/0.98	0.52/0.94	0.95	0.94
		Sparsistency (number of extra variables)						
hsnr	p=14	6(0.4)/6(0.2)/6(0.6)	6(0.6)	6(0.6)	6(3.8)/6(4.6)	6(0.9)/6(1.4)	6(0.7)	6(0.7)
	p=30	6(0.1)/6(0)/6(0.6)	6(0.7)	6(0.6)	6(7.3)/6(8.4)	6(2.3)/6(1.7)	6(1)	6(0.8)
	p=60	6(0.1)/6(0)/6(0.5)	-	6(0.5)	6(10)/6(11.3)	6(4.6)/6(1.8)	6(1.4)	6(0.9)
	p=180	6(8.9)/6(0)/6(0.4)	-	6(0.4)	6(15.3)/6(20.3)	6(18.1)/6(2.1)	6(2.3)	6(0.9)
msnr	p=14	6(0.8)/6(0.6)/6(0.6)	6(0.6)	6(0.6)	6(3.8)/6(4.6)	6(1.1)/6(1.4)	6(0.6)	6(0.8)
	p=30	6(0.4)/6(0.2)/6(0.6)	6(0.7)	6(0.6)	6(7.4)/6(8.4)	6(2.9)/6(1.6)	6(0.7)	6(0.8)
	p=60	6(0.2)/6(0.1)/6(0.5)	-	6(0.5)	6(10)/6(11.3)	6(6.2)/6(1.6)	6(1)	6(0.9)
	p=180	6(9.1)/6(0.1)/6(0.4)	-	6(0.4)	6(15.3)/6(20.3)	6(27.7)/6(1.7)	6(1.4)	6(1)
lsnr	p=14	5.4(4.5)/5.2(4.1)/4.6(1.7)	4.5(1.6)	4.6(1.7)	5.5(3.3)/5.5(3.9)	5(1.5)/5.1(2.9)	5(2.7)	4.8(2)
	p=30	4(4.7)/3.1(2.1)/3.7(2)	3.7(2)	3.7(2)	5.3(6.4)/5.3(7.1)	4.9(3.8)/4.8(4.9)	4.8(5.3)	4.5(3.3)
	p=60	2.8(1.9)/2(0.3)/2.9(1.4)	-	3(1.4)	5.1(8.5)/5.2(9.2)	4.8(8.1)/4.5(6.3)	4.5(7)	4.2(3.9)
	p=180	2.1(13.8)/1(0.1)/1.8(0.8)	-	1.8(0.8)	4.4(11.8)/4.5(15.4)	4.7(36.6)/3.8(10.1)	4.1(11.9)	3.7(7.2)

Table S14: The performance of BOSS compared to other methods, Sparse-Ex1, ρ=0, n=2000

		BOSS C_p-hdf/AICc-hdf/CV	BS CV	FS CV	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BOSS						
hsnr	p=14	6/6/17	18	17	40/40	12/14	11	17
	p=30	3/3/20	21	21	72/69	19/16	12	23
	p=60	2/2/23	-	22	100/96	30/19	15	25
	p=180	1/1/21	-	21	136/132	53/20	14	25
msnr	p=14	6/6/17	18	17	41/40	14/17	12	17
	p=30	3/3/20	21	21	72/69	26/18	12	23
	p=60	2/2/23	-	22	100/96	46/22	15	25
	p=180	1/1/21	-	21	136/132	97/22	13	25
lsnr	p=14	8/8/17	18	17	41/40	21/20	12	17
	p=30	5/4/20	21	21	72/69	44/23	12	23
	p=60	5/4/23	-	22	100/96	84/27	16	25
	p=180	5/5/21	-	21	136/132	192/27	13	25
		Relative efficiency						
hsnr	p=14	1/1/0.9	0.9	0.91	0.76/0.76	0.95/0.92	0.95	0.9
	p=30	1/1/0.86	0.85	0.85	0.6/0.61	0.87/0.88	0.92	0.83
	p=60	1/1/0.83	-	0.83	0.51/0.52	0.78/0.85	0.89	0.82
	p=180	1/1/0.84	-	0.84	0.43/0.44	0.66/0.85	0.89	0.81
msnr	p=14	1/1/0.9	0.9	0.91	0.75/0.76	0.93/0.91	0.95	0.9
	p=30	1/1/0.86	0.85	0.85	0.6/0.61	0.82/0.87	0.92	0.83
	p=60	1/1/0.83	-	0.83	0.51/0.52	0.7/0.84	0.89	0.82
	p=180	1/1/0.83	-	0.83	0.43/0.44	0.51/0.83	0.9	0.81
lsnr	p=14	1/1/0.93	0.92	0.93	0.77/0.77	0.9/0.9	0.97	0.92
	p=30	1/1/0.87	0.86	0.87	0.61/0.62	0.73/0.85	0.93	0.85
	p=60	1/1/0.85	-	0.85	0.52/0.53	0.57/0.82	0.9	0.84
	p=180	1/1/0.86	-	0.86	0.44/0.45	0.36/0.83	0.93	0.84
		Sparsistency (number of extra variables)						
hsnr	p=14	6(0.2)/6(0.2)/6(0.6)	6(0.6)	6(0.6)	6(3.8)/6(4.4)	6(0.9)/6(1.2)	6(0.6)	6(0.6)
	p=30	6(0.1)/6(0.1)/6(0.5)	6(0.6)	6(0.6)	6(8.5)/6(8.8)	6(1.9)/6(1.7)	6(1)	6(0.8)
	p=60	6(0)/6(0)/6(0.5)	-	6(0.5)	6(13.4)/6(12.6)	6(4)/6(2.2)	6(1.6)	6(0.8)
	p=180	6(0)/6(0)/6(0.4)	-	6(0.3)	6(23)/6(20.1)	6(10.4)/6(2.5)	6(2.1)	6(0.6)
msnr	p=14	6(0.2)/6(0.2)/6(0.6)	6(0.6)	6(0.6)	6(3.8)/6(4.4)	6(1)/6(1.3)	6(0.6)	6(0.6)
	p=30	6(0.1)/6(0.1)/6(0.5)	6(0.6)	6(0.6)	6(8.5)/6(8.7)	6(2.2)/6(1.6)	6(0.9)	6(0.8)
	p=60	6(0)/6(0)/6(0.5)	-	6(0.5)	6(13.3)/6(12.6)	6(5)/6(1.9)	6(1.5)	6(0.8)
	p=180	6(0)/6(0)/6(0.4)	-	6(0.3)	6(22.8)/6(20)	6(15.7)/6(1.9)	6(1.8)	6(0.6)
lsnr	p=14	6(0.4)/6(0.4)/6(0.6)	6(0.6)	6(0.6)	6(3.8)/6(4.4)	6(1.1)/6(1.3)	6(0.5)	6(0.6)
	p=30	6(0.1)/6(0.1)/6(0.5)	6(0.6)	6(0.6)	6(8.5)/6(8.7)	6(2.9)/6(1.4)	6(0.7)	6(0.8)
	p=60	6(0.1)/6(0.1)/6(0.5)	-	6(0.5)	6(13.4)/6(12.5)	6(6.7)/6(1.6)	6(1)	6(0.8)
	p=180	6(0.1)/6(0.1)/6(0.4)	-	6(0.3)	6(23)/6(20)	6(23.6)/6(1.1)	6(1)	6(0.6)

Table S15: The performance of BOSS compared to other methods, Sparse-Ex1, ρ=0.5, n=200

		BOSS C_p-hdf/AICc-hdf/CV	BS CV	FS CV	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BOSS						
hsnr	p=14	7/5/20	18	21	39/39	16/19	13	24
	p=30	4/2/20	22	21	66/65	34/21	15	21
	p=60	3/2/21	-	23	92/89	57/25	16	24
	p=180	65/1/19	-	22	139/134	136/25	16	21
msnr	p=14	19/17/20	15	20	34/34	21/20	12	23
	p=30	13/9/22	23	23	66/65	50/26	17	26
	p=60	12/8/22	-	25	90/88	85/30	17	33
	p=180	46/9/21	-	25	126/125	211/31	18	44
lsnr	p=14	19/22/23	23	21	-2/-2	12/6	7	6
	p=30	26/27/25	24	23	-4/-5	9/2	2	2
	p=60	24/26/22	-	20	-4/-6	22/1	0	0
	p=180	48/12/14	-	14	-2/-2	91/2	1	3
		Relative efficiency						
hsnr	p=14	0.98/1/0.87	0.89	0.87	0.75/0.75	0.9/0.88	0.93	0.84
	p=30	0.99/1/0.85	0.84	0.84	0.61/0.62	0.77/0.85	0.89	0.84
	p=60	0.99/1/0.85	-	0.83	0.53/0.54	0.65/0.82	0.88	0.82
	p=180	0.61/1/0.85	-	0.83	0.42/0.43	0.43/0.81	0.87	0.83
msnr	p=14	0.95/0.96/0.94	0.98	0.93	0.84/0.84	0.93/0.94	1	0.92
	p=30	0.97/1/0.9	0.89	0.89	0.66/0.66	0.73/0.87	0.93	0.87
	p=60	0.97/1/0.88	-	0.87	0.57/0.58	0.59/0.83	0.93	0.82
	p=180	0.75/1/0.9	-	0.88	0.48/0.49	0.35/0.83	0.93	0.76
lsnr	p=14	0.82/0.8/0.8	0.8	0.81	1/1	0.88/0.92	0.91	0.92
	p=30	0.75/0.75/0.76	0.77	0.77	0.99/1	0.87/0.93	0.93	0.93
	p=60	0.76/0.75/0.77	-	0.78	0.99/1	0.77/0.94	0.95	0.94
	p=180	0.66/0.87/0.86	-	0.86	1/1	0.51/0.96	0.97	0.95
		Sparsistency (number of extra variables)						
hsnr	p=14	6(0.3)/6(0.2)/6(0.8)	6(0.6)	6(0.8)	6(3.8)/6(4.2)	6(0.9)/6(1.3)	6(0.6)	6(1.6)
	p=30	6(0.1)/6(0)/6(0.6)	6(0.7)	6(0.6)	6(6)/6(8.6)	6(2.5)/6(1.5)	6(1.1)	6(1)
	p=60	6(0)/6(0)/6(0.5)	-	6(0.6)	6(8.8)/6(12.5)	6(4.9)/6(1.7)	6(1.4)	6(1.1)
	p=180	6(9.4)/6(0)/6(0.3)	-	6(0.4)	6(11.7)/6(21.3)	6(16.6)/6(1.7)	6(1.9)	6(0.7)
msnr	p=14	6(1.2)/6(1)/6(1.1)	6(0.7)	6(1.1)	6(3.8)/6(4.2)	6(1.1)/6(1.5)	6(0.7)	6(1.9)
	p=30	6(0.4)/6(0.2)/6(0.7)	6(0.7)	6(0.7)	6(6.1)/6(8.6)	6(3.1)/6(1.6)	6(1)	6(1.3)
	p=60	6(0.2)/6(0.2)/6(0.6)	-	6(0.6)	6(8.8)/6(12.4)	6(6.3)/6(1.8)	6(1.1)	6(1.7)
	p=180	6(10.2)/6(0.1)/6(0.4)	-	6(0.4)	6(16)/6(21.1)	6(26.5)/6(1.6)	6(1.3)	6(1.8)
lsnr	p=14	4.4(3.5)/4.1(3.2)/3.8(2.4)	3.7(2)	3.8(2.2)	5(3.3)/5.1(3.5)	4(1.6)/4.6(2.8)	4.6(2.7)	4.4(2.5)
	p=30	3.7(4.4)/2.9(2.1)/3.4(2.7)	3.3(2.3)	3.4(2.3)	4.8(4.8)/5.2(7.4)	4.5(3.9)/4.7(5.2)	4.8(5.7)	4.5(4.2)
	p=60	2.4(2)/1.6(0.5)/2.5(1.8)	-	2.5(1.6)	4.4(6.4)/4.8(10)	4.4(8.2)/4.2(7)	4.4(7.9)	4.1(5.7)
	p=180	1.3(14.1)/0.3(0.1)/1(0.8)	-	1.1(0.7)	2.5(4.9)/3.2(12)	4.2(35.8)/2.9(9.2)	3(9.7)	2.6(7.3)

Table S16: The performance of BOSS compared to other methods, Sparse-Ex1, ρ=0.5, n=2000

		BOSS C_p-hdf/AICc-hdf/CV	BS CV	FS CV	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BOSS						
hsnr	p=14	7/6/17	19	17	38/39	12/14	12	17
	p=30	3/3/20	23	22	67/65	21/18	14	23
	p=60	2/2/21	-	21	91/89	29/19	15	25
	p=180	2/1/22	-	23	126/123	52/20	15	28
msnr	p=14	7/6/17	19	17	39/39	16/17	12	17
	p=30	3/3/20	23	22	67/66	29/20	14	23
	p=60	2/2/21	-	21	91/89	45/21	14	25
	p=180	2/1/22	-	23	126/123	95/22	15	27
lsnr	p=14	13/13/17	18	17	39/39	23/21	13	20
	p=30	6/6/20	23	22	67/65	49/25	15	23
	p=60	5/5/21	-	21	91/89	83/26	15	25
	p=180	4/4/22	-	23	126/123	192/28	15	27
		Relative efficiency						
hsnr	p=14	1/1/0.91	0.9	0.91	0.77/0.77	0.95/0.93	0.95	0.91
	p=30	1/1/0.86	0.84	0.84	0.62/0.62	0.85/0.87	0.9	0.84
	p=60	1/1/0.84	-	0.84	0.54/0.54	0.79/0.86	0.89	0.82
	p=180	1/1/0.83	-	0.83	0.45/0.45	0.67/0.85	0.88	0.8
msnr	p=14	1/1/0.91	0.9	0.91	0.77/0.76	0.92/0.91	0.95	0.91
	p=30	1/1/0.86	0.84	0.84	0.62/0.62	0.8/0.86	0.9	0.84
	p=60	1/1/0.84	-	0.84	0.54/0.54	0.71/0.84	0.89	0.82
	p=180	1/1/0.83	-	0.83	0.45/0.45	0.52/0.83	0.89	0.8
lsnr	p=14	1/1/0.96	0.95	0.96	0.81/0.81	0.92/0.93	1	0.94
	p=30	1/1/0.88	0.86	0.87	0.63/0.64	0.71/0.85	0.92	0.86
	p=60	1/1/0.87	-	0.87	0.55/0.56	0.57/0.84	0.91	0.84
	p=180	1/1/0.85	-	0.85	0.46/0.47	0.36/0.82	0.9	0.82
		Sparsistency (number of extra variables)						
hsnr	p=14	6(0.3)/6(0.3)/6(0.6)	6(0.6)	6(0.6)	6(3.8)/6(4.2)	6(0.9)/6(1)	6(0.6)	6(0.8)
	p=30	6(0.1)/6(0.1)/6(0.6)	6(0.7)	6(0.6)	6(6.6)/6(8.3)	6(2.1)/6(1.7)	6(1)	6(1)
	p=60	6(0)/6(0)/6(0.5)	-	6(0.4)	6(10.3)/6(12.1)	6(3.7)/6(2)	6(1.3)	6(0.9)
	p=180	6(0)/6(0)/6(0.5)	-	6(0.4)	6(13.8)/6(18.9)	6(9.3)/6(2.5)	6(2.3)	6(0.9)
msnr	p=14	6(0.3)/6(0.3)/6(0.6)	6(0.6)	6(0.6)	6(3.8)/6(4.2)	6(1)/6(1.2)	6(0.6)	6(0.8)
	p=30	6(0.1)/6(0.1)/6(0.6)	6(0.7)	6(0.6)	6(6.6)/6(8.3)	6(2.5)/6(1.6)	6(0.9)	6(1)
	p=60	6(0)/6(0)/6(0.5)	-	6(0.4)	6(10.4)/6(12)	6(4.6)/6(1.7)	6(1.2)	6(0.9)
	p=180	6(0)/6(0)/6(0.5)	-	6(0.4)	6(14)/6(18.7)	6(14.2)/6(2)	6(2.1)	6(0.9)
lsnr	p=14	6(0.6)/6(0.6)/6(0.6)	6(0.6)	6(0.6)	6(3.8)/6(4.3)	6(1.2)/6(1.3)	6(0.5)	6(1.1)
	p=30	6(0.2)/6(0.1)/6(0.6)	6(0.7)	6(0.6)	6(6.6)/6(8.4)	6(3.1)/6(1.5)	6(0.7)	6(1)
	p=60	6(0.1)/6(0.1)/6(0.5)	-	6(0.4)	6(10.3)/6(12)	6(6.5)/6(1.4)	6(0.8)	6(0.9)
	p=180	6(0.1)/6(0)/6(0.5)	-	6(0.4)	6(13.8)/6(18.8)	6(23)/6(1.4)	6(1.4)	6(0.9)

Table S17: The performance of BOSS compared to other methods, Sparse-Ex1, ρ=0.9, n=200

		BOSS C_p-hdf/AICc-hdf/CV	BS CV	FS CV	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BOSS						
hsnr	p=14	20/21/19	16	18	6/5	10/6	12	13
	p=30	16/16/28	15	28	24/24	26/16	12	25
	p=60	15/15/34	-	34	58/59	66/38	28	44
	p=180	59/8/35	-	36	98/98	153/45	27	35
msnr	p=14	26/27/18	17	16	-8/-9	11/1	3	0
	p=30	24/27/20	19	15	-11/-12	5/-5	-3	-6
	p=60	16/16/19	-	16	2/3	27/6	8	4
	p=180	26/7/15	-	16	31/31	111/17	15	10
lsnr	p=14	28/27/24	22	19	-9/-13	9/3	3	2
	p=30	19/18/21	19	15	-18/-21	5/-7	-9	-9
	p=60	17/18/20	-	14	-20/-21	6/-12	-14	-13
	p=180	47/21/18	-	14	-13/-14	53/-9	-10	-9
		Relative efficiency						
hsnr	p=14	0.87/0.87/0.89	0.91	0.89	0.99/1	0.95/0.99	0.94	0.93
	p=30	0.97/0.97/0.87	0.98	0.88	0.91/0.91	0.89/0.96	1	0.9
	p=60	1/0.99/0.85	-	0.85	0.73/0.72	0.69/0.83	0.9	0.8
	p=180	0.68/1/0.8	-	0.79	0.55/0.54	0.43/0.74	0.85	0.8
msnr	p=14	0.72/0.71/0.77	0.77	0.78	0.98/1	0.82/0.9	0.88	0.9
	p=30	0.71/0.69/0.74	0.74	0.76	0.99/1	0.84/0.93	0.91	0.93
	p=60	0.88/0.88/0.86	-	0.88	1/0.99	0.8/0.96	0.95	0.98
	p=180	0.86/1/0.93	-	0.92	0.82/0.82	0.51/0.92	0.93	0.98
lsnr	p=14	0.68/0.68/0.7	0.71	0.73	0.95/1	0.8/0.85	0.85	0.85
	p=30	0.67/0.67/0.66	0.67	0.69	0.96/1	0.76/0.85	0.87	0.87
	p=60	0.68/0.67/0.66	-	0.69	0.98/1	0.74/0.89	0.92	0.91
	p=180	0.59/0.71/0.73	-	0.76	1/1	0.57/0.95	0.96	0.95
		Sparsistency (number of extra variables)						
hsnr	p=14	5.6(2.8)/5.5(2.6)/5.7(2.8)	5.6(1.9)	5.6(2.6)	5.9(4)/6(4)	5.6(1.4)/5.8(2)	5.6(2.2)	5.9(4.4)
	p=30	5.6(1.5)/5.6(1.2)/5.8(3.2)	5.8(1.5)	5.8(2.8)	6(7.5)/6(8.6)	5.8(3.4)/5.8(3.4)	5.8(2.3)	6(6.5)
	p=60	5.9(1)/5.8(0.9)/5.9(2.8)	-	5.9(1.9)	6(10.2)/6(12.1)	5.9(6.6)/5.9(4)	5.9(3)	6(5.5)
	p=180	6(9.9)/5.9(0.4)/6(2.4)	-	6(1.2)	6(13.6)/6(18)	6(21.6)/6(4.8)	6(4.3)	6(3.2)
msnr	p=14	2.9(2)/2.7(1.7)/3.7(2.9)	3.5(2.4)	3.6(2.7)	4.9(3.7)/5(3.7)	3.3(1.6)/4.2(2.6)	4(2.6)	4.5(3.6)
	p=30	3.3(2.9)/3(2.1)/3.9(4.9)	3.6(3.6)	3.8(4.1)	5.1(7.3)/5.3(8.2)	4.1(3.8)/4.7(5.7)	4.6(5.5)	5(7.3)
	p=60	4.3(2.8)/4.1(2.1)/4.4(4.3)	-	4.4(3.2)	5.6(10.1)/5.6(11.8)	4.7(7.4)/5(6.9)	5(6.8)	5.4(8.7)
	p=180	5.1(11.7)/5(1.2)/5(2.3)	-	5(1.7)	5.9(14.8)/5.9(17.5)	5.3(28.6)/5.4(6.8)	5.4(6.3)	5.8(5.4)
lsnr	p=14	1(1)/0.9(0.9)/1.5(1.7)	1.4(1.4)	1.5(1.4)	2.6(2.5)/2.9(2.7)	1.5(1.4)/2(1.8)	2(1.9)	2.1(2.1)
	p=30	0.9(1.8)/0.8(1.4)/1.2(2.9)	1.1(2.4)	1.2(2.2)	2.2(4.8)/2.6(6)	1.6(3.2)/2(4.2)	2.1(4.5)	2.1(4.6)
	p=60	0.9(2)/0.7(1.4)/1.1(3.2)	-	1.1(2.4)	2.6(7)/2.9(9)	2(6.5)/2.3(6.6)	2.5(7.2)	2.4(6.9)
	p=180	1.3(14.7)/0.5(0.6)/1(2.6)	-	1.2(1.9)	2.8(9.7)/3.1(13.9)	2.6(26.4)/2.6(10)	2.8(11.4)	2.6(9)

Table S18: The performance of BOSS compared to other methods, Sparse-Ex1, ρ=0.9, n=2000

		BOSS C_p-hdf/AICc-hdf/CV	BS CV	FS CV	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BOSS						
hsnr	p=14	7/7/21	18	22	30/30	6/8	17	33
	p=30	4/3/22	25	23	58/57	14/12	21	38
	p=60	2/2/21	-	22	81/82	34/19	18	21
	p=180	1/1/18	-	22	114/113	68/18	15	22
msnr	p=14	15/15/21	14	21	22/20	14/12	14	23
	p=30	4/4/28	23	28	54/53	46/22	21	49
	p=60	2/2/22	-	24	81/82	76/24	20	30
	p=180	1/1/18	-	22	114/113	140/21	15	21
lsnr	p=14	27/28/17	17	15	-7/-9	10/2	4	0
	p=30	22/22/20	16	17	-7/-7	10/-1	0	-3
	p=60	9/9/17	-	16	15/15	42/13	16	14
	p=180	3/3/13	-	15	59/58	146/21	33	20
		Relative efficiency						
hsnr	p=14	0.99/0.99/0.87	0.89	0.87	0.81/0.81	1/0.97	0.9	0.8
	p=30	1/1/0.85	0.83	0.84	0.65/0.66	0.9/0.92	0.86	0.75
	p=60	1/1/0.84	-	0.83	0.56/0.56	0.76/0.86	0.86	0.84
	p=180	1/1/0.85	-	0.83	0.47/0.47	0.6/0.85	0.88	0.83
msnr	p=14	0.98/0.98/0.93	0.99	0.93	0.92/0.93	0.99/1	0.98	0.91
	p=30	1/1/0.81	0.85	0.81	0.67/0.68	0.71/0.85	0.86	0.7
	p=60	1/1/0.83	-	0.82	0.56/0.56	0.58/0.82	0.85	0.78
	p=180	1/1/0.85	-	0.83	0.47/0.47	0.42/0.84	0.87	0.83
lsnr	p=14	0.72/0.72/0.78	0.78	0.79	0.98/1	0.83/0.9	0.88	0.91
	p=30	0.76/0.76/0.77	0.8	0.79	0.99/1	0.84/0.94	0.92	0.95
	p=60	1/1/0.93	-	0.93	0.94/0.94	0.77/0.96	0.94	0.95
	p=180	1/1/0.92	-	0.9	0.65/0.65	0.42/0.85	0.78	0.86
		Sparsistency (number of extra variables)						
hsnr	p=14	6(0.3)/6(0.3)/6(0.9)	6(0.7)	6(0.9)	6(3.4)/6(3.4)	6(0.2)/6(0.3)	6(0.6)	6(3.1)
	p=30	6(0.1)/6(0.1)/6(0.7)	6(0.7)	6(0.6)	6(7.6)/6(8)	6(0.9)/6(0.7)	6(1.2)	6(2.8)
	p=60	6(0)/6(0)/6(1)	-	6(0.5)	6(11.1)/6(12.1)	6(2.9)/6(1.4)	6(1.4)	6(0.9)
	p=180	6(0)/6(0)/6(1.3)	-	6(0.4)	6(15.6)/6(18.5)	6(9.5)/6(1.7)	6(1.9)	6(0.8)
msnr	p=14	5.9(1.2)/5.9(1.2)/6(1.7)	6(1)	6(1.6)	6(4)/6(3.9)	6(1)/6(1.2)	5.9(1)	6(3.9)
	p=30	6(0.2)/6(0.2)/6(1.2)	6(0.7)	6(1)	6(7.8)/6(8.2)	6(3.1)/6(1.4)	6(1.2)	6(4.9)
	p=60	6(0)/6(0)/6(1.1)	-	6(0.6)	6(11.1)/6(12.1)	6(6.4)/6(1.5)	6(1.2)	6(1.4)
	p=180	6(0)/6(0)/6(1.3)	-	6(0.4)	6(15.8)/6(18.5)	6(18.7)/6(1.4)	6(1.5)	6(0.7)
lsnr	p=14	3.5(2.1)/3.5(2.1)/4.3(3)	4(2.4)	4.2(2.8)	5.3(3.8)/5.4(3.8)	3.9(1.8)/4.6(2.7)	4.5(2.6)	5(3.7)
	p=30	4.1(2.7)/4.1(2.7)/4.6(4.5)	4.5(3.2)	4.5(3.7)	5.6(7.7)/5.7(8)	4.7(4)/5.1(5.3)	5.1(5)	5.5(7.4)
	p=60	5(1.4)/5(1.4)/5.1(3.1)	-	5.1(2.2)	5.9(11.1)/5.9(12)	5.3(8.2)/5.4(6)	5.3(5.7)	5.8(7.8)
	p=180	5.6(0.6)/5.6(0.6)/5.6(2)	-	5.6(0.9)	6(15.7)/6(18.4)	5.7(26.1)/5.6(3.6)	5.5(5.3)	5.9(3.4)

Table S19: The performance of BOSS compared to other methods, Sparse-Ex2, ρ=0, n=200

		BOSS C_p-hdf/AICc-hdf/CV	BS CV	FS CV	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BOSS						
hsnr	p=14	8/6/20	21	20	41/41	17/20	15	20
	p=30	5/3/24	25	25	69/68	32/22	15	24
	p=60	4/2/21	-	23	95/94	53/23	16	26
	p=180	34/1/19	-	21	129/130	139/27	18	25
msnr	p=14	17/14/20	21	20	42/41	23/23	16	20
	p=30	17/13/24	25	25	69/68	48/27	16	24
	p=60	13/9/21	-	23	95/94	84/28	16	29
	p=180	49/10/20	-	22	129/130	224/32	18	35
lsnr	p=14	21/22/24	25	24	7/7	14/14	16	14
	p=30	29/31/26	25	26	1/1	13/8	8	8
	p=60	26/28/23	-	22	0/0	25/5	6	7
	p=180	31/18/15	-	16	-2/0	85/4	3	5
		Relative efficiency						
hsnr	p=14	0.98/1/0.89	0.88	0.89	0.75/0.75	0.91/0.89	0.92	0.89
	p=30	0.98/1/0.83	0.82	0.82	0.61/0.61	0.78/0.84	0.9	0.83
	p=60	0.99/1/0.85	-	0.83	0.52/0.53	0.67/0.83	0.88	0.81
	p=180	0.75/1/0.85	-	0.84	0.44/0.44	0.42/0.8	0.86	0.81
msnr	p=14	0.98/1/0.95	0.94	0.95	0.8/0.81	0.93/0.93	0.98	0.95
	p=30	0.96/1/0.91	0.9	0.9	0.67/0.67	0.76/0.89	0.97	0.91
	p=60	0.97/1/0.9	-	0.89	0.56/0.56	0.59/0.85	0.94	0.85
	p=180	0.74/1/0.92	-	0.9	0.48/0.48	0.34/0.83	0.93	0.81
lsnr	p=14	0.89/0.88/0.86	0.86	0.86	1/1	0.94/0.94	0.92	0.94
	p=30	0.78/0.77/0.8	0.8	0.8	1/1	0.89/0.93	0.93	0.93
	p=60	0.79/0.78/0.81	-	0.81	1/0.99	0.8/0.95	0.94	0.93
	p=180	0.75/0.83/0.85	-	0.85	1/0.98	0.53/0.94	0.95	0.93
		Sparsistency (number of extra variables)						
hsnr	p=14	6(0.3)/6(0.2)/6(0.6)	6(0.7)	6(0.6)	6(3.6)/6(4.4)	6(1)/6(1.4)	6(0.6)	6(0.7)
	p=30	6(0.1)/6(0)/6(0.6)	6(0.7)	6(0.7)	6(7.5)/6(8.4)	6(2.4)/6(1.7)	6(1.1)	6(0.9)
	p=60	6(0.1)/6(0)/6(0.5)	-	6(0.5)	6(11.4)/6(13.2)	6(4.8)/6(1.6)	6(1.6)	6(0.9)
	p=180	6(8.8)/6(0)/6(0.4)	-	6(0.4)	6(15.9)/6(22.1)	6(18.5)/6(2.2)	6(2.6)	6(0.9)
msnr	p=14	6(0.8)/6(0.6)/6(0.6)	6(0.7)	6(0.6)	6(3.6)/6(4.4)	6(1.1)/6(1.4)	6(0.6)	6(0.7)
	p=30	6(0.5)/6(0.2)/6(0.6)	6(0.7)	6(0.7)	6(7.4)/6(8.5)	6(2.9)/6(1.6)	6(0.8)	6(0.9)
	p=60	6(0.2)/6(0.1)/6(0.5)	-	6(0.5)	6(11.3)/6(13.2)	6(6.4)/6(1.5)	6(1)	6(1.1)
	p=180	6(9.8)/6(0.1)/6(0.4)	-	6(0.4)	6(15.9)/6(22)	6(27.5)/6(2)	6(1.6)	6(1.2)
lsnr	p=14	5.5(4.5)/5.4(4.2)/4.8(1.8)	4.8(1.7)	4.8(1.8)	5.7(3.3)/5.7(4)	5.1(1.5)/5.3(2.9)	5.2(2.8)	5.1(1.8)
	p=30	3.9(4.3)/3.1(1.8)/3.6(2.2)	3.6(2)	3.6(2.1)	5.3(6.6)/5.3(7.1)	4.8(3.8)/4.7(4.9)	4.8(5.3)	4.5(3.5)
	p=60	2.3(1.8)/1.4(0.3)/2.7(1.4)	-	2.7(1.4)	4.7(8.8)/4.7(9.9)	4.9(8.6)/4.2(6.9)	4.3(7.5)	3.9(5.2)
	p=180	1.7(14.1)/0.5(0.1)/1.4(0.6)	-	1.4(0.7)	3.7(10.7)/3.8(13.5)	4.6(36.7)/3.2(9.4)	3.3(10.3)	3(7.1)

Table S20: The performance of BOSS compared to other methods, Sparse-Ex2, ρ=0, n=2000

		BOSS C_p-hdf/AICc-hdf/CV	BS CV	FS CV	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BOSS						
hsnr	p=14	6/6/17	17	17	40/41	12/14	11	17
	p=30	3/3/22	22	23	74/71	20/18	13	24
	p=60	2/2/24	-	23	97/93	29/19	15	26
	p=180	1/1/21	-	21	131/128	52/18	13	24
msnr	p=14	6/6/17	17	17	41/41	14/17	11	17
	p=30	3/3/22	22	23	74/71	27/20	13	24
	p=60	2/2/24	-	23	97/93	44/21	15	26
	p=180	1/1/21	-	21	131/127	97/22	13	24
lsnr	p=14	9/8/17	17	17	41/41	21/21	12	17
	p=30	5/5/22	22	23	74/71	46/25	13	24
	p=60	5/4/24	-	23	97/93	82/26	15	26
	p=180	5/5/21	-	21	131/128	192/26	12	24
		Relative efficiency						
hsnr	p=14	1/1/0.91	0.91	0.91	0.76/0.76	0.95/0.93	0.96	0.91
	p=30	1/1/0.84	0.84	0.84	0.59/0.6	0.86/0.87	0.91	0.83
	p=60	1/1/0.83	-	0.83	0.52/0.53	0.79/0.86	0.89	0.81
	p=180	1/1/0.83	-	0.84	0.44/0.44	0.66/0.86	0.89	0.81
msnr	p=14	1/1/0.91	0.91	0.91	0.75/0.75	0.93/0.91	0.96	0.91
	p=30	1/1/0.84	0.84	0.84	0.59/0.6	0.81/0.86	0.91	0.83
	p=60	1/1/0.83	-	0.83	0.52/0.53	0.71/0.84	0.89	0.81
	p=180	1/1/0.83	-	0.84	0.44/0.44	0.51/0.83	0.9	0.81
lsnr	p=14	1/1/0.93	0.93	0.93	0.77/0.77	0.9/0.9	0.97	0.93
	p=30	1/1/0.86	0.86	0.85	0.6/0.61	0.72/0.84	0.93	0.84
	p=60	1/1/0.84	-	0.85	0.53/0.54	0.57/0.83	0.91	0.83
	p=180	1/1/0.86	-	0.86	0.45/0.46	0.36/0.83	0.94	0.84
		Sparsistency (number of extra variables)						
hsnr	p=14	6(0.3)/6(0.3)/6(0.6)	6(0.6)	6(0.6)	6(3.8)/6(4.5)	6(0.9)/6(1.3)	6(0.6)	6(0.7)
	p=30	6(0.1)/6(0.1)/6(0.6)	6(0.6)	6(0.6)	6(8.4)/6(8.7)	6(2)/6(1.8)	6(0.9)	6(0.9)
	p=60	6(0)/6(0)/6(0.5)	-	6(0.5)	6(13.1)/6(12.2)	6(3.8)/6(2.1)	6(1.5)	6(0.8)
	p=180	6(0)/6(0)/6(0.3)	-	6(0.3)	6(21.6)/6(19.3)	6(10.3)/6(2.4)	6(2.1)	6(0.6)
msnr	p=14	6(0.3)/6(0.3)/6(0.6)	6(0.6)	6(0.6)	6(3.8)/6(4.5)	6(0.9)/6(1.3)	6(0.6)	6(0.7)
	p=30	6(0.1)/6(0.1)/6(0.6)	6(0.6)	6(0.6)	6(8.5)/6(8.7)	6(2.2)/6(1.7)	6(0.9)	6(0.9)
	p=60	6(0)/6(0)/6(0.5)	-	6(0.5)	6(13.1)/6(12.2)	6(4.8)/6(1.8)	6(1.4)	6(0.8)
	p=180	6(0)/6(0)/6(0.3)	-	6(0.3)	6(21.8)/6(19.4)	6(15.6)/6(1.9)	6(1.9)	6(0.6)
lsnr	p=14	6(0.4)/6(0.4)/6(0.6)	6(0.6)	6(0.6)	6(3.8)/6(4.5)	6(1.1)/6(1.4)	6(0.5)	6(0.7)
	p=30	6(0.1)/6(0.1)/6(0.6)	6(0.6)	6(0.6)	6(8.4)/6(8.7)	6(3)/6(1.6)	6(0.7)	6(0.9)
	p=60	6(0.1)/6(0.1)/6(0.5)	-	6(0.5)	6(13.1)/6(12.2)	6(6.8)/6(1.4)	6(1)	6(0.8)
	p=180	6(0.1)/6(0.1)/6(0.3)	-	6(0.3)	6(21.7)/6(19.4)	6(23.9)/6(1.1)	6(0.9)	6(0.6)

Table S21: The performance of BOSS compared to other methods, Sparse-Ex2, ρ=0.5, n=200

		BOSS C_p-hdf/AICc-hdf/CV	BS CV	FS CV	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BOSS						
hsnr	p=14	8/6/17	18	17	46/44	15/20	13	18
	p=30	5/3/24	24	24	91/88	28/23	14	24
	p=60	4/2/22	-	23	123/119	43/24	16	24
	p=180	34/1/19	-	22	168/165	100/28	17	22
msnr	p=14	19/16/26	18	17	46/44	21/23	14	18
	p=30	22/18/29	23	25	89/87	42/28	14	49
	p=60	16/11/29	-	25	121/117	70/29	16	55
	p=180	48/9/29	-	30	160/157	179/32	14	71
lsnr	p=14	24/25/28	23	26	18/17	15/20	22	24
	p=30	33/34/26	20	30	17/17	15/19	18	25
	p=60	28/29/23	-	28	15/16	23/17	16	21
	p=180	28/14/14	-	17	8/9	71/11	10	14
		Relative efficiency						
hsnr	p=14	0.98/1/0.91	0.9	0.91	0.73/0.74	0.92/0.89	0.94	0.9
	p=30	0.98/1/0.83	0.83	0.83	0.54/0.55	0.8/0.84	0.9	0.83
	p=60	0.98/1/0.83	-	0.83	0.46/0.46	0.71/0.82	0.88	0.82
	p=180	0.75/1/0.85	-	0.83	0.38/0.38	0.5/0.79	0.86	0.83
msnr	p=14	0.96/0.98/0.9	0.97	0.97	0.78/0.79	0.94/0.92	1	0.96
	p=30	0.93/0.97/0.88	0.92	0.91	0.6/0.61	0.8/0.89	1	0.76
	p=60	0.96/1/0.86	-	0.89	0.5/0.51	0.66/0.87	0.96	0.72
	p=180	0.74/1/0.84	-	0.83	0.42/0.42	0.39/0.83	0.95	0.63
lsnr	p=14	0.92/0.92/0.9	0.94	0.91	0.97/0.98	1/0.96	0.94	0.93
	p=30	0.87/0.86/0.92	0.96	0.88	0.99/0.98	1/0.96	0.98	0.92
	p=60	0.89/0.89/0.93	-	0.89	1/0.99	0.93/0.98	0.99	0.94
	p=180	0.84/0.94/0.94	-	0.92	1/0.99	0.63/0.97	0.98	0.95
		Sparsistency (number of extra variables)						
hsnr	p=14	6(0.4)/6(0.3)/6(0.6)	6(0.6)	6(0.6)	6(4.6)/6(5.5)	6(0.9)/6(1.5)	6(0.7)	6(0.7)
	p=30	6(0.1)/6(0)/6(0.7)	6(0.7)	6(0.7)	6(11.2)/6(12.8)	6(2.2)/6(1.5)	6(1.1)	6(0.9)
	p=60	6(0.1)/6(0)/6(0.5)	-	6(0.5)	6(16)/6(19.1)	6(4.3)/6(1.5)	6(1.7)	6(0.9)
	p=180	6(9.3)/6(0)/6(0.4)	-	6(0.4)	6(22.5)/6(31.5)	6(14.2)/6(2.1)	6(2.8)	6(0.7)
msnr	p=14	6(0.9)/6(0.8)/6(0.6)	6(0.6)	6(0.6)	6(4.6)/6(5.5)	6(1.1)/6(1.6)	6(0.7)	6(0.7)
	p=30	6(0.6)/6(0.4)/6(0.7)	6(0.7)	6(0.7)	6(11.2)/6(12.8)	6(2.7)/6(1.5)	6(0.9)	6(2.5)
	p=60	6(0.3)/6(0.2)/6(0.5)	-	6(0.6)	6(16)/6(19.1)	6(5.6)/6(1.3)	6(1.4)	6(2.5)
	p=180	6(10.1)/6(0.1)/6(0.5)	-	6(0.7)	6(22.3)/6(31.6)	6(23.9)/6(2)	6(2.1)	6(3.4)
lsnr	p=14	5.7(4.8)/5.6(4.5)/5.1(1.6)	5.1(1.4)	5.1(1.9)	5.7(4.3)/5.7(5.1)	5.4(1.6)/5.5(3.7)	5.3(2.7)	5.4(3.1)
	p=30	3.6(5.3)/2.4(2.2)/3.9(2.4)	3.9(1.9)	3.5(2.8)	4.5(8)/4.4(8.9)	5(4.7)/4.3(6.5)	4.2(5.3)	3.8(6.1)
	p=60	2.3(2.4)/1(0.3)/3.1(2.1)	-	2.5(1.8)	3.8(9.8)/3.7(10.9)	5(9)/3.7(8.1)	3.7(7)	3.2(7.5)
	p=180	1.7(14.9)/0.4(0.1)/1.4(1.1)	-	0.9(0.8)	2.4(9.8)/2.4(12.9)	4.5(35.8)/2.3(11)	2.3(9.3)	2.1(9.1)

Table S22: The performance of BOSS compared to other methods, Sparse-Ex2, ρ=0.5, n=2000

		BOSS C_p-hdf/AICc-hdf/CV	BS CV	FS CV	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BOSS						
hsnr	p=14	7/7/18	19	18	48/47	14/18	13	18
	p=30	4/3/23	22	23	86/83	20/18	12	23
	p=60	2/2/23	-	23	124/120	28/20	14	24
	p=180	1/1/21	-	21	174/171	46/20	14	24
msnr	p=14	7/7/18	19	18	49/48	16/19	13	18
	p=30	4/3/23	22	23	85/83	25/20	12	23
	p=60	2/2/23	-	23	125/120	39/22	14	24
	p=180	1/1/21	-	21	174/171	77/23	13	24
lsnr	p=14	12/12/23	19	18	50/48	21/23	13	18
	p=30	9/8/27	22	23	85/83	40/25	13	23
	p=60	8/8/26	-	23	124/120	72/28	14	24
	p=180	10/9/24	-	21	175/170	163/29	12	25
		Relative efficiency						
hsnr	p=14	1/1/0.91	0.9	0.91	0.72/0.72	0.94/0.91	0.94	0.9
	p=30	1/1/0.84	0.85	0.84	0.56/0.56	0.86/0.87	0.92	0.84
	p=60	1/1/0.83	-	0.83	0.45/0.46	0.79/0.85	0.89	0.82
	p=180	1/1/0.84	-	0.84	0.37/0.37	0.69/0.84	0.89	0.81
msnr	p=14	1/1/0.91	0.9	0.91	0.72/0.72	0.92/0.89	0.94	0.9
	p=30	1/1/0.84	0.85	0.84	0.56/0.56	0.83/0.86	0.92	0.84
	p=60	1/1/0.83	-	0.83	0.45/0.46	0.73/0.83	0.89	0.82
	p=180	1/1/0.84	-	0.84	0.37/0.37	0.57/0.82	0.89	0.81
lsnr	p=14	1/1/0.91	0.94	0.95	0.75/0.76	0.92/0.91	0.99	0.95
	p=30	1/1/0.85	0.89	0.88	0.58/0.59	0.77/0.87	0.96	0.88
	p=60	1/1/0.85	-	0.88	0.48/0.49	0.63/0.84	0.95	0.87
	p=180	0.99/1/0.88	-	0.9	0.4/0.4	0.41/0.85	0.97	0.87
		Sparsistency (number of extra variables)						
hsnr	p=14	6(0.3)/6(0.3)/6(0.6)	6(0.7)	6(0.6)	6(4.7)/6(5.5)	6(1)/6(1.4)	6(0.7)	6(0.7)
	p=30	6(0.1)/6(0.1)/6(0.6)	6(0.6)	6(0.6)	6(10.9)/6(11.3)	6(2.1)/6(1.9)	6(0.9)	6(0.7)
	p=60	6(0)/6(0)/6(0.5)	-	6(0.5)	6(18.1)/6(17.7)	6(3.8)/6(2.1)	6(1.4)	6(0.7)
	p=180	6(0)/6(0)/6(0.3)	-	6(0.3)	6(32.2)/6(29.7)	6(9.2)/6(2.4)	6(2.1)	6(0.5)
msnr	p=14	6(0.3)/6(0.3)/6(0.6)	6(0.7)	6(0.6)	6(4.8)/6(5.5)	6(1)/6(1.5)	6(0.6)	6(0.7)
	p=30	6(0.1)/6(0.1)/6(0.6)	6(0.6)	6(0.6)	6(11)/6(11.3)	6(2.3)/6(1.7)	6(0.9)	6(0.7)
	p=60	6(0)/6(0)/6(0.5)	-	6(0.5)	6(18.1)/6(17.7)	6(4.6)/6(1.8)	6(1.4)	6(0.7)
	p=180	6(0)/6(0)/6(0.3)	-	6(0.3)	6(32.3)/6(29.7)	6(13.4)/6(1.8)	6(2)	6(0.5)
lsnr	p=14	6(0.6)/6(0.5)/6(0.6)	6(0.7)	6(0.6)	6(4.8)/6(5.5)	6(1.1)/6(1.5)	6(0.6)	6(0.7)
	p=30	6(0.2)/6(0.2)/6(0.6)	6(0.6)	6(0.6)	6(10.9)/6(11.3)	6(2.8)/6(1.5)	6(0.8)	6(0.7)
	p=60	6(0.1)/6(0.1)/6(0.5)	-	6(0.5)	6(18.1)/6(17.8)	6(6.4)/6(1.4)	6(1.1)	6(0.7)
	p=180	6(0.1)/6(0.1)/6(0.3)	-	6(0.3)	6(32.5)/6(29.7)	6(21.7)/6(1.1)	6(1.2)	6(0.5)

Table S23: The performance of BOSS compared to other methods, Sparse-Ex2, ρ=0.9, n=200

		BOSS C_p-hdf/AICc-hdf/CV	BS CV	FS CV	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BOSS						
hsnr	p=14	18/15/25	17	23	53/51	20/27	14	29
	p=30	12/9/42	9	53	91/89	18/18	1	57
	p=60	9/6/49	-	69	131/126	16/12	-6	74
	p=180	22/5/75	-	111	138/120	-7/-16	-30	70
msnr	p=14	25/23/29	3	24	35/34	10/16	3	27
	p=30	23/20/31	-15	36	50/48	-1/2	-17	40
	p=60	21/18/30	-	48	65/61	-7/-7	-26	52
	p=180	23/22/20	-	64	48/35	-16/-1	-24	35
lsnr	p=14	40/41/31	32	50	41/40	38/39	39	46
	p=30	42/42/28	28	61	53/53	44/51	46	57
	p=60	36/35/25	-	50	45/45	52/47	45	49
	p=180	25/9/11	-	15	10/11	59/12	12	14
		Relative efficiency						
hsnr	p=14	0.97/0.99/0.91	0.97	0.93	0.75/0.75	0.95/0.89	1	0.89
	p=30	0.9/0.92/0.71	0.93	0.66	0.53/0.53	0.86/0.85	1	0.64
	p=60	0.87/0.89/0.63	-	0.56	0.41/0.42	0.81/0.84	1	0.54
	p=180	0.58/0.67/0.4	-	0.33	0.3/0.32	0.76/0.84	1	0.41
msnr	p=14	0.82/0.83/0.79	0.99	0.83	0.76/0.77	0.94/0.88	1	0.81
	p=30	0.68/0.69/0.63	0.98	0.61	0.55/0.56	0.84/0.82	1	0.59
	p=60	0.61/0.62/0.57	-	0.5	0.45/0.46	0.79/0.79	1	0.49
	p=180	0.62/0.62/0.63	-	0.46	0.51/0.56	0.9/0.76	1	0.56
lsnr	p=14	0.94/0.93/1	0.99	0.87	0.92/0.93	0.95/0.94	0.94	0.9
	p=30	0.9/0.9/1	1	0.8	0.83/0.84	0.89/0.85	0.88	0.81
	p=60	0.92/0.93/1	-	0.83	0.86/0.86	0.83/0.85	0.87	0.84
	p=180	0.87/1/0.98	-	0.95	0.98/0.98	0.68/0.97	0.97	0.96
		Sparsistency (number of extra variables)						
hsnr	p=14	6(0.9)/6(0.7)/6(0.7)	6(0.6)	6(1)	6(6.5)/6(7.2)	6(1.3)/6(2.2)	6(0.7)	6(1.6)
	p=30	6(1.1)/6(0.9)/6(2.5)	6(0.7)	6(3.7)	6(18.3)/6(19.9)	6(2.9)/6(2.5)	6(1.1)	6(7.6)
	p=60	6(1.7)/6(1.6)/6(3.3)	-	6(4.9)	6(31.8)/6(36.4)	6(4.2)/6(2.7)	6(1.8)	6(10.6)
	p=180	6(14.9)/6(6.8)/6(11.8)	-	5.9(19.7)	6(51.3)/6(76.1)	6(8.6)/6(4.1)	6(3.2)	6(23.4)
msnr	p=14	6(3)/6(2.7)/6(1.2)	6(0.6)	6(2.4)	6(6.5)/6(7.2)	6(1.5)/6(3)	6(1.3)	6(3.2)
	p=30	6(5.1)/6(4)/6(3.5)	6(0.7)	6(7)	6(18.2)/6(19.9)	6(3.6)/6(5.3)	6(3.4)	6(11.7)
	p=60	6(6.1)/6(4.7)/6(5.1)	-	6(10.3)	6(31.8)/6(36.3)	6(6.5)/6(7.7)	6(6)	6(19.4)
	p=180	5.9(38.2)/5.5(14.8)/5.8(18.2)	-	4.2(16.8)	5.8(49.6)/6(75.8)	6(23.5)/5.9(36.2)	5.9(31.7)	5.9(49.1)
lsnr	p=14	5.7(4.9)/5.6(4.6)/5.6(2.7)	5.2(1.2)	5.3(4)	5.8(6.3)/5.8(6.9)	5.4(3.7)/5.7(5.9)	5.4(4.7)	5.6(5.8)
	p=30	4(7.5)/2.9(4.4)/4.6(6.6)	4.1(1.7)	2.5(4.5)	2.7(8.3)/2.8(9.2)	4.4(9.7)/3.3(9.4)	3.4(8.7)	2.5(7.7)
	p=60	2.6(7.1)/1.4(2.5)/3.8(9.8)	-	0.8(1.7)	0.8(4.4)/0.9(5.7)	3.7(16)/1.3(7.2)	1.8(9.1)	0.8(4.6)
	p=180	1.2(20.1)/0.3(0.5)/0.7(2.7)	-	0.1(0.5)	0.3(4.4)/0.3(4.6)	2(35.8)/0.3(4.5)	0.3(4.9)	0.2(3.2)

Table S24: The performance of BOSS compared to other methods, Sparse-Ex2, ρ=0.9, n=2000

		BOSS C_p-hdf/AICc-hdf/CV	BS CV	FS CV	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BOSS						
hsnr	p=14	7/7/18	19	18	56/55	22/28	14	18
	p=30	4/3/23	22	23	118/116	29/29	14	22
	p=60	2/2/23	-	23	186/182	37/33	15	20
	p=180	1/1/21	-	21	299/294	51/41	15	20
msnr	p=14	8/8/18	19	18	56/55	23/29	14	18
	p=30	4/4/24	22	23	118/115	31/30	13	22
	p=60	3/3/23	-	23	185/182	42/36	16	20
	p=180	3/3/22	-	21	298/293	61/43	14	20
lsnr	p=14	36/36/42	15	19	52/50	22/30	12	28
	p=30	37/36/43	12	32	100/97	28/26	7	49
	p=60	39/38/47	-	47	141/138	35/22	1	75
	p=180	38/38/37	-	72	178/175	45/16	-12	112
		Relative efficiency						
hsnr	p=14	1/1/0.91	0.9	0.91	0.68/0.69	0.88/0.84	0.94	0.9
	p=30	1/1/0.84	0.85	0.84	0.47/0.48	0.8/0.8	0.9	0.84
	p=60	1/1/0.83	-	0.83	0.35/0.36	0.74/0.76	0.88	0.84
	p=180	1/1/0.84	-	0.84	0.25/0.26	0.67/0.72	0.88	0.84
msnr	p=14	1/1/0.91	0.91	0.92	0.69/0.7	0.88/0.83	0.94	0.91
	p=30	1/1/0.84	0.85	0.84	0.48/0.48	0.79/0.8	0.91	0.85
	p=60	1/1/0.84	-	0.84	0.36/0.36	0.72/0.76	0.88	0.85
	p=180	1/1/0.84	-	0.85	0.26/0.26	0.64/0.72	0.9	0.85
lsnr	p=14	0.83/0.83/0.79	0.98	0.94	0.74/0.75	0.92/0.87	1	0.88
	p=30	0.79/0.79/0.75	0.96	0.81	0.54/0.54	0.84/0.85	1	0.72
	p=60	0.72/0.73/0.68	-	0.69	0.42/0.42	0.75/0.83	1	0.57
	p=180	0.64/0.64/0.64	-	0.51	0.32/0.32	0.61/0.76	1	0.42
		Sparsistency (number of extra variables)						
hsnr	p=14	6(0.3)/6(0.3)/6(0.6)	6(0.7)	6(0.6)	6(6.6)/6(7.2)	6(1.5)/6(2.1)	6(0.7)	6(0.7)
	p=30	6(0.1)/6(0.1)/6(0.6)	6(0.6)	6(0.6)	6(17.8)/6(18.9)	6(2.7)/6(2.6)	6(0.9)	6(0.6)
	p=60	6(0)/6(0)/6(0.5)	-	6(0.5)	6(34.2)/6(35.3)	6(4.5)/6(2.9)	6(1.4)	6(0.5)
	p=180	6(0)/6(0)/6(0.3)	-	6(0.3)	6(72.8)/6(73.6)	6(8.7)/6(3.9)	6(2.2)	6(0.4)
msnr	p=14	6(0.3)/6(0.3)/6(0.6)	6(0.7)	6(0.6)	6(6.6)/6(7.2)	6(1.5)/6(2.2)	6(0.7)	6(0.7)
	p=30	6(0.1)/6(0.1)/6(0.6)	6(0.6)	6(0.6)	6(17.8)/6(18.8)	6(2.8)/6(2.3)	6(0.9)	6(0.6)
	p=60	6(0)/6(0)/6(0.5)	-	6(0.5)	6(34.2)/6(35.2)	6(4.8)/6(2.6)	6(1.4)	6(0.5)
	p=180	6(0)/6(0)/6(0.3)	-	6(0.3)	6(72.8)/6(73.6)	6(10.5)/6(3.3)	6(2.1)	6(0.4)
lsnr	p=14	6(2.7)/6(2.6)/6(0.6)	6(0.7)	6(0.9)	6(6.6)/6(7.2)	6(1.5)/6(2.5)	6(0.8)	6(1.6)
	p=30	6(2.1)/6(2)/6(0.8)	6(0.6)	6(1.5)	6(17.8)/6(18.8)	6(3.2)/6(2.7)	6(1.4)	6(2.9)
	p=60	6(1)/6(0.9)/6(0.9)	-	6(2.2)	6(34.2)/6(35.1)	6(6.3)/6(3.7)	6(2.6)	6(5)
	p=180	6(1.4)/6(1.4)/6(1.8)	-	6(4.6)	6(72.3)/6(73.5)	6(17.1)/6(8.4)	6(7.1)	6(11.9)

Table S25: The performance of BOSS compared to other methods, Sparse-Ex3, ρ=0, n=200

		BOSS C_p-hdf/AICc-hdf/CV	BS CV	FS CV	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BOSS						
hsnr	p=14	8/6/20	21	20	44/43	17/20	15	20
	p=30	5/3/24	25	25	69/67	32/23	15	25
	p=60	4/2/21	-	23	97/96	52/24	16	25
	p=180	34/1/19	-	21	133/133	137/28	19	23
msnr	p=14	17/14/20	21	20	44/43	24/24	16	20
	p=30	18/13/24	25	25	69/67	48/27	16	25
	p=60	14/9/21	-	23	97/95	84/29	16	28
	p=180	50/11/20	-	22	132/133	224/33	19	36
lsnr	p=14	22/23/26	26	26	8/8	13/15	17	16
	p=30	29/32/26	26	25	1/1	14/8	8	8
	p=60	27/29/22	-	22	0/1	24/6	6	7
	p=180	30/16/14	-	14	-2/1	84/4	3	6
		Relative efficiency						
hsnr	p=14	0.98/1/0.89	0.88	0.89	0.74/0.75	0.91/0.88	0.93	0.89
	p=30	0.98/1/0.83	0.82	0.82	0.61/0.61	0.78/0.84	0.89	0.83
	p=60	0.99/1/0.85	-	0.83	0.52/0.52	0.67/0.83	0.88	0.82
	p=180	0.75/1/0.85	-	0.84	0.43/0.43	0.43/0.79	0.85	0.82
msnr	p=14	0.98/1/0.95	0.95	0.96	0.8/0.8	0.92/0.92	0.99	0.95
	p=30	0.96/1/0.92	0.91	0.91	0.67/0.68	0.76/0.89	0.98	0.9
	p=60	0.96/1/0.9	-	0.89	0.56/0.56	0.59/0.85	0.94	0.85
	p=180	0.74/1/0.92	-	0.91	0.48/0.47	0.34/0.83	0.93	0.81
lsnr	p=14	0.89/0.88/0.86	0.86	0.86	1/1	0.95/0.94	0.92	0.93
	p=30	0.78/0.76/0.8	0.8	0.81	1/1	0.88/0.93	0.93	0.93
	p=60	0.79/0.78/0.82	-	0.82	1/1	0.81/0.95	0.94	0.93
	p=180	0.76/0.85/0.86	-	0.86	1/0.98	0.53/0.94	0.95	0.93
		Sparsistency (number of extra variables)						
hsnr	p=14	6(0.3)/6(0.2)/6(0.6)	6(0.7)	6(0.6)	6(3.7)/6(4.5)	6(1)/6(1.3)	6(0.6)	6(0.8)
	p=30	6(0.1)/6(0)/6(0.6)	6(0.7)	6(0.7)	6(7.4)/6(8.2)	6(2.4)/6(1.7)	6(1)	6(0.9)
	p=60	6(0.1)/6(0)/6(0.5)	-	6(0.5)	6(11.3)/6(13.1)	6(4.8)/6(1.7)	6(1.5)	6(0.9)
	p=180	6(8.8)/6(0)/6(0.4)	-	6(0.4)	6(16.5)/6(22.7)	6(18)/6(2.5)	6(2.7)	6(0.7)
msnr	p=14	6(0.8)/6(0.6)/6(0.6)	6(0.7)	6(0.6)	6(3.7)/6(4.5)	6(1.2)/6(1.4)	6(0.6)	6(0.8)
	p=30	6(0.5)/6(0.3)/6(0.6)	6(0.7)	6(0.7)	6(7.4)/6(8.2)	6(2.9)/6(1.6)	6(0.8)	6(0.9)
	p=60	6(0.3)/6(0.1)/6(0.5)	-	6(0.5)	6(11.4)/6(13.1)	6(6.4)/6(1.5)	6(1.1)	6(1)
	p=180	6(9.4)/6(0.1)/6(0.4)	-	6(0.4)	6(16.5)/6(22.8)	6(27.4)/6(2)	6(1.7)	6(1.3)
lsnr	p=14	5.4(4.4)/5.2(4.1)/4.7(1.8)	4.7(1.7)	4.7(1.7)	5.6(3.3)/5.6(4)	5.1(1.5)/5.3(3)	5.1(2.8)	5(1.9)
	p=30	4(4.4)/3.1(1.9)/3.7(2.1)	3.6(2.1)	3.7(2)	5.3(6.4)/5.4(7)	4.9(3.8)/4.8(4.9)	4.8(5.3)	4.5(3.3)
	p=60	2.2(1.8)/1.2(0.2)/2.6(1.4)	-	2.6(1.3)	4.6(8.6)/4.6(9.6)	4.9(8.5)/4.1(6.6)	4.2(7.2)	3.9(5)
	p=180	1.6(14.2)/0.5(0.1)/1.3(0.6)	-	1.3(0.6)	3.4(10.4)/3.5(13.1)	4.6(36.9)/3(9)	3.2(10.5)	2.8(7.1)

Table S26: The performance of BOSS compared to other methods, Sparse-Ex3, ρ=0, n=2000

		BOSS C_p-hdf/AICc-hdf/CV	BS CV	FS CV	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BOSS						
hsnr	p=14	6/6/17	17	17	41/41	12/15	11	17
	p=30	3/3/22	22	23	72/69	20/18	13	24
	p=60	2/2/24	-	23	97/93	29/19	14	24
	p=180	1/1/21	-	21	132/129	53/19	13	24
msnr	p=14	6/6/17	17	17	41/41	14/17	11	17
	p=30	3/3/22	22	23	72/69	26/20	13	24
	p=60	2/2/24	-	23	97/93	43/21	14	24
	p=180	1/1/21	-	21	132/129	97/22	13	24
lsnr	p=14	9/9/17	17	17	42/41	21/20	12	17
	p=30	5/5/22	22	23	72/69	45/25	13	24
	p=60	5/4/24	-	23	97/93	82/26	15	24
	p=180	5/4/21	-	21	132/129	192/26	13	24
		Relative efficiency						
hsnr	p=14	1/1/0.91	0.91	0.91	0.76/0.76	0.95/0.93	0.96	0.91
	p=30	1/1/0.84	0.84	0.84	0.6/0.61	0.86/0.87	0.91	0.83
	p=60	1/1/0.83	-	0.83	0.52/0.53	0.79/0.86	0.89	0.82
	p=180	1/1/0.83	-	0.84	0.44/0.44	0.66/0.85	0.89	0.81
msnr	p=14	1/1/0.91	0.91	0.91	0.75/0.75	0.93/0.91	0.96	0.91
	p=30	1/1/0.84	0.84	0.84	0.6/0.61	0.81/0.86	0.91	0.83
	p=60	1/1/0.83	-	0.83	0.52/0.53	0.71/0.84	0.89	0.82
	p=180	1/1/0.83	-	0.84	0.44/0.44	0.51/0.83	0.89	0.81
lsnr	p=14	1/1/0.93	0.93	0.93	0.77/0.77	0.9/0.9	0.97	0.93
	p=30	0.99/1/0.86	0.86	0.85	0.61/0.62	0.72/0.84	0.93	0.84
	p=60	1/1/0.84	-	0.85	0.53/0.54	0.57/0.83	0.91	0.84
	p=180	0.99/1/0.86	-	0.86	0.45/0.46	0.36/0.82	0.92	0.84
		Sparsistency (number of extra variables)						
hsnr	p=14	6(0.3)/6(0.3)/6(0.6)	6(0.6)	6(0.6)	6(3.8)/6(4.5)	6(0.9)/6(1.3)	6(0.6)	6(0.7)
	p=30	6(0.1)/6(0.1)/6(0.6)	6(0.6)	6(0.6)	6(8.3)/6(8.5)	6(2)/6(1.8)	6(0.9)	6(0.8)
	p=60	6(0)/6(0)/6(0.5)	-	6(0.5)	6(13)/6(12.2)	6(3.9)/6(2.1)	6(1.5)	6(0.7)
	p=180	6(0)/6(0)/6(0.3)	-	6(0.3)	6(21.7)/6(19.1)	6(10.3)/6(2.4)	6(2.1)	6(0.6)
msnr	p=14	6(0.3)/6(0.3)/6(0.6)	6(0.6)	6(0.6)	6(3.9)/6(4.5)	6(1)/6(1.3)	6(0.6)	6(0.6)
	p=30	6(0.1)/6(0.1)/6(0.6)	6(0.6)	6(0.6)	6(8.3)/6(8.6)	6(2.2)/6(1.8)	6(0.9)	6(0.8)
	p=60	6(0)/6(0)/6(0.5)	-	6(0.5)	6(13)/6(12.2)	6(4.8)/6(1.8)	6(1.4)	6(0.7)
	p=180	6(0)/6(0)/6(0.3)	-	6(0.3)	6(21.6)/6(19.1)	6(15.6)/6(1.8)	6(1.8)	6(0.6)
lsnr	p=14	6(0.4)/6(0.4)/6(0.6)	6(0.6)	6(0.6)	6(3.9)/6(4.5)	6(1.1)/6(1.4)	6(0.5)	6(0.7)
	p=30	6(0.1)/6(0.1)/6(0.6)	6(0.6)	6(0.6)	6(8.3)/6(8.5)	6(3)/6(1.6)	6(0.7)	6(0.8)
	p=60	6(0.1)/6(0.1)/6(0.5)	-	6(0.5)	6(13.2)/6(12.2)	6(6.8)/6(1.5)	6(1)	6(0.7)
	p=180	6(0.1)/6(0)/6(0.3)	-	6(0.3)	6(21.9)/6(19.1)	6(23.7)/6(1)	6(1)	6(0.6)

Table S27: The performance of BOSS compared to other methods, Sparse-Ex3, ρ=0.5, n=200

		BOSS C_p-hdf/AICc-hdf/CV	BS CV	FS CV	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BOSS						
hsnr	p=14	7/6/18	19	18	40/39	15/18	14	18
	p=30	5/2/22	24	22	70/68	30/21	14	22
	p=60	3/1/22	-	23	93/92	51/22	16	24
	p=180	35/1/18	-	21	135/135	134/26	17	24
msnr	p=14	15/13/18	19	18	40/39	20/21	15	19
	p=30	14/10/22	25	23	69/68	45/25	15	25
	p=60	13/9/23	-	24	91/89	78/26	15	33
	p=180	48/11/20	-	23	132/133	218/31	18	51
lsnr	p=14	19/21/24	24	24	5/4	11/11	13	12
	p=30	28/30/25	25	25	0/0	12/6	7	8
	p=60	23/25/21	-	21	-3/-3	22/3	3	4
	p=180	27/11/13	-	13	-3/-1	83/3	3	4
		Relative efficiency						
hsnr	p=14	0.98/1/0.9	0.89	0.9	0.75/0.76	0.92/0.9	0.92	0.9
	p=30	0.98/1/0.84	0.82	0.84	0.6/0.61	0.79/0.85	0.9	0.84
	p=60	0.98/1/0.83	-	0.83	0.53/0.53	0.67/0.83	0.88	0.82
	p=180	0.75/1/0.85	-	0.84	0.43/0.43	0.43/0.8	0.86	0.81
msnr	p=14	0.98/1/0.95	0.95	0.95	0.8/0.81	0.93/0.93	0.98	0.94
	p=30	0.97/1/0.9	0.88	0.89	0.65/0.66	0.76/0.88	0.96	0.88
	p=60	0.97/1/0.88	-	0.88	0.57/0.57	0.61/0.86	0.94	0.82
	p=180	0.75/1/0.92	-	0.9	0.48/0.47	0.35/0.84	0.94	0.73
lsnr	p=14	0.88/0.86/0.84	0.84	0.84	0.99/1	0.94/0.94	0.93	0.93
	p=30	0.78/0.77/0.8	0.8	0.8	1/1	0.9/0.94	0.93	0.92
	p=60	0.79/0.78/0.8	-	0.8	1/1	0.79/0.94	0.94	0.93
	p=180	0.76/0.87/0.86	-	0.86	1/0.98	0.53/0.94	0.94	0.93
		Sparsistency (number of extra variables)						
hsnr	p=14	6(0.4)/6(0.2)/6(0.6)	6(0.7)	6(0.6)	6(3.7)/6(4.5)	6(0.9)/6(1.3)	6(0.7)	6(0.8)
	p=30	6(0.2)/6(0)/6(0.6)	6(0.7)	6(0.6)	6(7.9)/6(9)	6(2.4)/6(1.5)	6(1.1)	6(1)
	p=60	6(0.1)/6(0)/6(0.5)	-	6(0.5)	6(11.5)/6(13.3)	6(4.9)/6(1.6)	6(1.6)	6(1)
	p=180	6(9.6)/6(0)/6(0.3)	-	6(0.4)	6(16.6)/6(23.8)	6(18.1)/6(2.1)	6(2.4)	6(1)
msnr	p=14	6(0.7)/6(0.6)/6(0.6)	6(0.7)	6(0.6)	6(3.7)/6(4.5)	6(1)/6(1.3)	6(0.6)	6(1)
	p=30	6(0.4)/6(0.2)/6(0.7)	6(0.8)	6(0.7)	6(8)/6(9)	6(2.9)/6(1.5)	6(0.8)	6(1.1)
	p=60	6(0.3)/6(0.2)/6(0.6)	-	6(0.6)	6(11.5)/6(13.3)	6(6.1)/6(1.5)	6(1.1)	6(1.7)
	p=180	6(10)/6(0.1)/6(0.4)	-	6(0.5)	6(17)/6(23.8)	6(27.9)/6(1.8)	6(1.7)	6(2.3)
lsnr	p=14	5.3(4.3)/5.2(4)/4.6(2.1)	4.6(2)	4.6(2)	5.5(3.4)/5.6(4)	4.9(1.4)/5.2(3)	5.2(2.6)	5(2.3)
	p=30	3.8(4.3)/2.9(2)/3.6(2.4)	3.4(2.1)	3.5(2.3)	5.1(6.9)/5.2(7.5)	4.6(3.9)/4.6(5.3)	4.7(5.3)	4.4(4)
	p=60	2.1(1.7)/1.3(0.4)/2.3(1.5)	-	2.3(1.4)	4.4(8.6)/4.5(9.7)	4.4(8.3)/3.9(6.7)	4(7.5)	3.7(5.2)
	p=180	1.4(15.1)/0.3(0.1)/1(0.7)	-	1(0.7)	3(9.7)/3(12.5)	4.3(37.2)/2.5(8.3)	2.6(9.3)	2.3(7.1)

Table S28: The performance of BOSS compared to other methods, Sparse-Ex3, ρ=0.5, n=2000

		BOSS C_p-hdf/AICc-hdf/CV	BS CV	FS CV	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BOSS						
hsnr	p=14	7/6/18	19	19	43/42	13/16	13	18
	p=30	3/3/22	22	21	73/71	21/18	14	23
	p=60	1/1/21	-	20	96/92	29/18	13	26
	p=180	1/1/22	-	22	130/126	53/20	14	25
msnr	p=14	7/6/18	19	19	44/43	16/18	13	18
	p=30	3/3/22	22	21	73/71	28/20	14	23
	p=60	1/1/21	-	20	96/92	43/21	13	26
	p=180	1/1/22	-	22	130/126	98/22	13	25
lsnr	p=14	9/9/18	19	19	44/43	22/22	13	18
	p=30	5/5/22	21	21	73/71	47/25	14	23
	p=60	4/4/21	-	20	96/92	80/26	14	26
	p=180	5/5/22	-	22	129/126	192/27	13	24
		Relative efficiency						
hsnr	p=14	1/1/0.9	0.9	0.9	0.75/0.75	0.94/0.92	0.95	0.9
	p=30	1/1/0.85	0.85	0.85	0.59/0.6	0.85/0.87	0.91	0.84
	p=60	1/1/0.84	-	0.84	0.52/0.53	0.79/0.86	0.89	0.8
	p=180	1/1/0.83	-	0.83	0.44/0.45	0.66/0.84	0.89	0.81
msnr	p=14	1/1/0.9	0.9	0.9	0.74/0.75	0.92/0.9	0.94	0.9
	p=30	1/1/0.85	0.85	0.85	0.59/0.6	0.8/0.85	0.9	0.84
	p=60	1/1/0.84	-	0.84	0.52/0.53	0.71/0.84	0.9	0.8
	p=180	1/1/0.83	-	0.83	0.44/0.45	0.51/0.83	0.89	0.81
lsnr	p=14	1/1/0.92	0.92	0.92	0.75/0.76	0.89/0.89	0.96	0.92
	p=30	1/1/0.86	0.86	0.86	0.61/0.61	0.71/0.84	0.92	0.85
	p=60	1/1/0.86	-	0.86	0.53/0.54	0.58/0.83	0.91	0.82
	p=180	1/1/0.86	-	0.86	0.46/0.46	0.36/0.82	0.93	0.84
		Sparsistency (number of extra variables)						
hsnr	p=14	6(0.3)/6(0.3)/6(0.6)	6(0.6)	6(0.6)	6(3.9)/6(4.4)	6(0.9)/6(1.3)	6(0.6)	6(0.8)
	p=30	6(0.1)/6(0.1)/6(0.6)	6(0.6)	6(0.6)	6(8.4)/6(8.6)	6(2.1)/6(1.7)	6(1)	6(0.8)
	p=60	6(0)/6(0)/6(0.5)	-	6(0.5)	6(13.2)/6(12.3)	6(3.9)/6(2.1)	6(1.6)	6(0.9)
	p=180	6(0)/6(0)/6(0.4)	-	6(0.4)	6(21.5)/6(18.7)	6(10.4)/6(2.6)	6(2.3)	6(0.7)
msnr	p=14	6(0.3)/6(0.3)/6(0.6)	6(0.6)	6(0.6)	6(3.9)/6(4.5)	6(1)/6(1.4)	6(0.6)	6(0.8)
	p=30	6(0.1)/6(0.1)/6(0.6)	6(0.6)	6(0.6)	6(8.3)/6(8.6)	6(2.4)/6(1.6)	6(0.9)	6(0.8)
	p=60	6(0)/6(0)/6(0.5)	-	6(0.5)	6(13)/6(12.3)	6(4.7)/6(1.8)	6(1.4)	6(0.9)
	p=180	6(0)/6(0)/6(0.4)	-	6(0.4)	6(21.5)/6(18.7)	6(15.7)/6(1.9)	6(2.1)	6(0.7)
lsnr	p=14	6(0.4)/6(0.4)/6(0.6)	6(0.6)	6(0.6)	6(3.9)/6(4.5)	6(1.1)/6(1.4)	6(0.5)	6(0.8)
	p=30	6(0.1)/6(0.1)/6(0.6)	6(0.6)	6(0.6)	6(8.3)/6(8.6)	6(3)/6(1.4)	6(0.7)	6(0.8)
	p=60	6(0.1)/6(0.1)/6(0.5)	-	6(0.5)	6(13.1)/6(12.2)	6(6.5)/6(1.5)	6(1)	6(0.9)
	p=180	6(0.1)/6(0.1)/6(0.4)	-	6(0.4)	6(21.2)/6(18.7)	6(23.5)/6(1.2)	6(0.9)	6(0.7)

Table S29: The performance of BOSS compared to other methods, Sparse-Ex3, ρ=0.9, n=200

		BOSS C_p-hdf/AICc-hdf/CV	BS CV	FS CV	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BOSS						
hsnr	p=14	7/6/24	13	24	33/33	14/16	12	20
	p=30	7/5/41	17	41	66/64	26/29	12	33
	p=60	6/4/43	-	43	84/83	44/38	13	39
	p=180	35/3/27	-	29	126/126	132/35	16	30
msnr	p=14	14/13/19	16	19	18/17	17/17	6	11
	p=30	15/13/24	8	25	30/29	29/24	0	12
	p=60	12/10/20	-	20	43/42	50/25	-6	14
	p=180	35/7/16	-	18	87/87	164/23	2	21
lsnr	p=14	17/20/22	22	21	-2/-3	5/3	5	5
	p=30	26/28/24	23	23	-2/-2	8/6	5	5
	p=60	23/26/22	-	22	-3/-3	20/3	4	3
	p=180	29/16/16	-	15	-5/-3	79/1	1	2
		Relative efficiency						
hsnr	p=14	0.99/1/0.85	0.93	0.85	0.79/0.79	0.92/0.91	0.95	0.88
	p=30	0.98/1/0.74	0.9	0.74	0.63/0.64	0.83/0.81	0.93	0.79
	p=60	0.98/1/0.73	-	0.73	0.57/0.57	0.73/0.75	0.92	0.75
	p=180	0.76/1/0.8	-	0.79	0.45/0.45	0.44/0.76	0.88	0.79
msnr	p=14	0.93/0.94/0.9	0.92	0.9	0.9/0.91	0.91/0.91	1	0.96
	p=30	0.87/0.88/0.8	0.93	0.8	0.77/0.77	0.77/0.8	1	0.89
	p=60	0.84/0.85/0.78	-	0.78	0.66/0.66	0.63/0.75	1	0.83
	p=180	0.75/0.95/0.88	-	0.87	0.55/0.55	0.39/0.83	1	0.84
lsnr	p=14	0.83/0.81/0.8	0.8	0.81	0.99/1	0.93/0.94	0.93	0.93
	p=30	0.78/0.76/0.79	0.8	0.8	1/1	0.91/0.93	0.93	0.94
	p=60	0.78/0.77/0.79	-	0.8	1/1	0.8/0.94	0.93	0.94
	p=180	0.74/0.82/0.82	-	0.83	1/0.98	0.53/0.94	0.95	0.93
		Sparsistency (number of extra variables)						
hsnr	p=14	6(0.6)/6(0.5)/6(1.4)	6(0.6)	6(1.4)	6(3.9)/6(4.5)	6(1)/6(1.4)	6(0.7)	6(1.8)
	p=30	6(0.7)/6(0.6)/6(2.1)	6(0.8)	6(2.1)	6(9.2)/6(10.3)	6(2.5)/6(3.4)	6(1.6)	6(3.2)
	p=60	6(0.8)/6(0.7)/6(2)	-	6(1.9)	6(12.4)/6(14.2)	6(5.1)/6(4.7)	6(2.5)	6(3.1)
	p=180	6(9.2)/6(0.1)/6(0.6)	-	6(0.6)	6(16.2)/6(22.2)	6(18.1)/6(3.4)	6(2.4)	6(1.7)
msnr	p=14	5.6(1.2)/5.6(1.1)/5.8(2.1)	5.8(1.8)	5.8(2.1)	6(4)/6(4.5)	5.7(1.5)/5.8(2.6)	5.8(1)	5.9(2.3)
	p=30	5.1(1.5)/5.1(1.2)/5.3(2.6)	5.5(1.4)	5.3(2.6)	6(9.1)/6(10.2)	5.2(4.1)/5.4(5.8)	5.7(1.8)	5.8(3.8)
	p=60	5.2(1.2)/5.2(1)/5.2(1.7)	-	5.2(1.6)	5.9(12.4)/6(14.1)	5.2(7.5)/5.3(5.2)	5.8(1.7)	5.8(3.4)
	p=180	5.6(10)/5.6(0.5)/5.6(0.8)	-	5.6(0.8)	6(16.2)/6(22.2)	5.6(27.5)/5.6(3.1)	5.9(1.7)	5.9(2.6)
lsnr	p=14	4.4(4)/4.2(3.6)/3.7(2.5)	3.6(2.4)	3.7(2.3)	4.8(3.6)/4.9(4.1)	3.9(2.1)/4.4(3.2)	4.5(2.5)	4.4(3)
	p=30	2.6(4.4)/1.9(2.3)/2.4(3)	2.5(2.8)	2.4(2.9)	3.9(7.5)/4(8.2)	3.3(4.9)/3.3(6.3)	3.7(6.1)	3.3(5.1)
	p=60	1.7(2)/1.1(0.8)/1.8(2)	-	1.8(1.9)	3.7(9.4)/3.8(10.4)	3.4(8.9)/3.2(7.6)	3.4(7.7)	3(6)
	p=180	1.4(14.4)/0.5(0.2)/1.1(1.1)	-	1.1(1.1)	3.2(11.1)/3.3(14.7)	3.5(36.7)/2.8(10.2)	3(10.8)	2.6(7.5)

Table S30: The performance of BOSS compared to other methods, Sparse-Ex3, ρ=0.9, n=2000

		BOSS C_p-hdf/AICc-hdf/CV	BS CV	FS CV	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BOSS						
hsnr	p=14	6/6/19	19	19	40/39	12/14	13	17
	p=30	2/2/21	21	22	74/72	20/18	13	22
	p=60	1/1/22	-	22	101/97	29/19	14	25
	p=180	1/1/21	-	22	135/131	52/18	14	25
msnr	p=14	6/6/19	19	19	42/40	18/18	14	17
	p=30	2/2/21	21	22	74/72	27/20	14	22
	p=60	1/1/22	-	22	101/97	43/21	14	25
	p=180	1/1/21	-	22	135/132	96/21	13	26
lsnr	p=14	9/9/19	19	19	27/26	17/18	6	15
	p=30	5/5/21	20	21	53/51	34/23	3	18
	p=60	4/4/20	-	20	75/71	62/22	1	17
	p=180	4/4/17	-	17	92/89	145/23	-5	16
		Relative efficiency						
hsnr	p=14	1/1/0.89	0.89	0.89	0.76/0.76	0.95/0.93	0.94	0.9
	p=30	1/1/0.84	0.85	0.84	0.59/0.6	0.85/0.87	0.9	0.84
	p=60	1/1/0.83	-	0.83	0.5/0.51	0.79/0.85	0.89	0.81
	p=180	1/1/0.84	-	0.83	0.43/0.44	0.67/0.86	0.89	0.81
msnr	p=14	1/1/0.89	0.89	0.89	0.75/0.76	0.9/0.9	0.93	0.9
	p=30	1/1/0.84	0.85	0.84	0.59/0.6	0.81/0.85	0.9	0.84
	p=60	1/1/0.83	-	0.83	0.5/0.51	0.71/0.84	0.89	0.81
	p=180	1/1/0.84	-	0.83	0.43/0.44	0.52/0.83	0.9	0.81
lsnr	p=14	0.97/0.97/0.89	0.89	0.89	0.83/0.84	0.91/0.89	1	0.92
	p=30	0.98/0.98/0.85	0.86	0.85	0.67/0.68	0.77/0.83	1	0.87
	p=60	0.97/0.97/0.84	-	0.84	0.58/0.59	0.62/0.82	1	0.86
	p=180	0.91/0.91/0.81	-	0.81	0.49/0.5	0.39/0.77	1	0.81
		Sparsistency (number of extra variables)						
hsnr	p=14	6(0.3)/6(0.3)/6(0.7)	6(0.6)	6(0.6)	6(4.1)/6(4.6)	6(0.8)/6(1)	6(0.7)	6(0.8)
	p=30	6(0)/6(0)/6(0.6)	6(0.6)	6(0.6)	6(9.2)/6(9.5)	6(2)/6(1.8)	6(1)	6(1)
	p=60	6(0)/6(0)/6(0.5)	-	6(0.5)	6(14)/6(13.3)	6(3.7)/6(2.1)	6(1.5)	6(1)
	p=180	6(0)/6(0)/6(0.4)	-	6(0.4)	6(23.2)/6(20.6)	6(10.1)/6(2.4)	6(2.2)	6(1)
msnr	p=14	6(0.3)/6(0.3)/6(0.7)	6(0.6)	6(0.6)	6(4.2)/6(4.6)	6(1.1)/6(1.3)	6(0.7)	6(0.8)
	p=30	6(0)/6(0)/6(0.6)	6(0.6)	6(0.6)	6(9.2)/6(9.5)	6(2.3)/6(1.7)	6(1)	6(1)
	p=60	6(0)/6(0)/6(0.5)	-	6(0.5)	6(14.1)/6(13.3)	6(4.7)/6(1.8)	6(1.4)	6(1)
	p=180	6(0)/6(0)/6(0.4)	-	6(0.4)	6(23.2)/6(20.5)	6(15.3)/6(1.9)	6(1.9)	6(1)
lsnr	p=14	5.8(0.4)/5.8(0.4)/5.9(1.4)	5.9(1.5)	5.9(1.5)	6(4.2)/6(4.6)	5.9(1.2)/5.9(2.1)	6(0.8)	6(1.7)
	p=30	5.8(0.3)/5.8(0.3)/5.8(1.1)	5.8(1.1)	5.8(1.1)	6(9.2)/6(9.5)	5.8(2.8)/5.9(3)	6(0.8)	6(1.9)
	p=60	5.8(0.3)/5.8(0.3)/5.8(0.8)	-	5.8(0.8)	6(14)/6(13.3)	5.8(6.3)/5.8(2.3)	6(1.1)	6(1.6)
	p=180	5.7(0.3)/5.7(0.3)/5.7(0.7)	-	5.7(0.7)	6(23)/6(20.6)	5.7(23.2)/5.7(2.6)	6(1)	5.9(2.1)

Table S31: The performance of BOSS compared to other methods, Sparse-Ex4, ρ=0, n=200

		BOSS C_p-hdf/AICc-hdf/CV	BS CV	FS CV	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BOSS						
hsnr	p=14	30/30/22	22	22	19/19	16/18	20	20
	p=30	24/21/24	25	25	29/28	32/19	16	21
	p=60	16/15/20	-	21	37/36	57/17	13	24
	p=180	29/5/14	-	15	50/51	168/14	11	19
msnr	p=14	25/22/22	23	22	31/31	26/21	18	21
	p=30	26/20/27	28	28	52/51	43/24	21	27
	p=60	18/14/24	-	25	69/68	72/25	21	31
	p=180	55/16/26	-	27	97/99	237/29	27	39
lsnr	p=14	34/34/29	30	29	16/16	19/25	25	19
	p=30	37/34/31	31	31	20/18	33/25	24	23
	p=60	32/33/29	-	29	20/20	55/24	22	23
	p=180	49/29/27	-	26	18/20	142/23	21	18
		Relative efficiency						
hsnr	p=14	0.89/0.89/0.95	0.95	0.95	0.98/0.98	1/0.98	0.97	0.97
	p=30	0.94/0.96/0.94	0.93	0.93	0.9/0.91	0.88/0.98	1	0.96
	p=60	0.97/0.99/0.94	-	0.94	0.83/0.83	0.72/0.97	1	0.92
	p=180	0.82/1/0.92	-	0.92	0.7/0.7	0.39/0.92	0.94	0.89
msnr	p=14	0.95/0.97/0.97	0.97	0.97	0.9/0.91	0.94/0.98	1	0.98
	p=30	0.96/1/0.95	0.94	0.94	0.79/0.8	0.84/0.97	0.99	0.95
	p=60	0.96/1/0.92	-	0.92	0.68/0.68	0.66/0.91	0.95	0.87
	p=180	0.74/1/0.92	-	0.91	0.59/0.58	0.34/0.9	0.91	0.83
lsnr	p=14	0.86/0.87/0.9	0.89	0.9	1/1	0.97/0.93	0.93	0.98
	p=30	0.87/0.89/0.9	0.91	0.9	0.99/1	0.89/0.95	0.96	0.97
	p=60	0.91/0.9/0.93	-	0.93	1/1	0.77/0.97	0.98	0.97
	p=180	0.79/0.91/0.93	-	0.93	1/0.98	0.49/0.95	0.97	1
		Sparsistency (number of extra variables)						
hsnr	p=14	5(1.7)/4.9(1.3)/5.2(0.9)	5.2(0.9)	5.2(0.9)	5.8(3.4)/5.9(4.2)	5.5(1.3)/5.3(1.5)	5.3(1.3)	5.3(1.1)
	p=30	4.6(0.7)/4.4(0.2)/4.9(0.9)	4.9(1.1)	4.9(1)	5.7(7)/5.8(8)	5.4(3.3)/5(1.9)	5(2)	5(1.5)
	p=60	4.3(0.2)/4.2(0)/4.6(0.7)	-	4.6(0.7)	5.5(10.4)/5.6(12.1)	5.4(7.5)/4.8(2.2)	4.9(2.8)	4.7(1.9)
	p=180	4.3(10.2)/4.1(0)/4.2(0.4)	-	4.3(0.5)	5.2(14)/5.3(19)	5.4(35.4)/4.4(2.3)	4.5(3.9)	4.3(1.7)
msnr	p=14	4.5(1.3)/4.4(1)/4.4(0.6)	4.4(0.7)	4.4(0.7)	5.2(3.1)/5.2(3.6)	4.6(1.1)/4.5(1)	4.4(0.8)	4.5(0.8)
	p=30	4.2(0.8)/4.1(0.4)/4.2(0.8)	4.2(0.8)	4.2(0.8)	5(6.5)/5(7.1)	4.6(2.9)/4.3(1.3)	4.3(1.3)	4.3(1.2)
	p=60	4(0.3)/4(0.2)/4.1(0.6)	-	4.1(0.6)	4.7(9.5)/4.7(10.7)	4.5(6.5)/4.1(1.3)	4.2(1.7)	4.1(1.4)
	p=180	4.1(10.4)/3.9(0.1)/4(0.5)	-	4(0.5)	4.4(13.1)/4.5(17.1)	4.6(31.6)/4(1.6)	4.1(2.6)	4(1.8)
lsnr	p=14	3.9(2.4)/3.7(2)/3.3(0.9)	3.3(1.1)	3.2(0.9)	4.4(2.7)/4.5(3)	3.7(1.3)/3.6(1.5)	3.6(1.6)	3.6(1.2)
	p=30	2.6(1.8)/2.2(0.7)/2.5(1.1)	2.5(1.1)	2.5(1.1)	3.8(5.3)/3.9(5.5)	3.4(3.5)/2.9(2.5)	3(2.9)	2.9(2.1)
	p=60	2(0.8)/1.7(0.2)/2.2(0.8)	-	2.2(0.8)	3.4(7.3)/3.5(8.1)	3.5(7.8)/2.7(3.3)	2.8(3.9)	2.7(2.9)
	p=180	1.8(12)/1.3(0.1)/1.6(0.6)	-	1.7(0.6)	2.9(9.7)/2.9(11.2)	3.7(34)/2.2(5)	2.4(6.2)	2.2(3.7)

Table S32: The performance of BOSS compared to other methods, Sparse-Ex4, ρ=0, n=2000

		BOSS C_p-hdf/AICc-hdf/CV	BS CV	FS CV	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BOSS						
hsnr	p=14	9/9/17	17	17	41/41	19/15	12	17
	p=30	5/5/22	22	23	74/71	45/22	13	24
	p=60	4/4/24	-	23	97/93	83/23	16	26
	p=180	4/4/21	-	21	131/127	200/23	13	24
msnr	p=14	31/31/21	21	22	34/33	20/20	22	21
	p=30	40/40/29	31	29	52/50	42/29	28	33
	p=60	40/40/32	-	32	64/61	76/29	29	33
	p=180	39/40/31	-	32	67/64	157/27	27	30
lsnr	p=14	19/19/19	19	19	28/28	24/18	16	19
	p=30	11/11/21	22	21	49/47	57/20	14	21
	p=60	10/9/20	-	20	64/61	102/19	13	23
	p=180	9/8/20	-	20	91/88	235/21	14	23
		Relative efficiency						
hsnr	p=14	1/1/0.93	0.93	0.93	0.77/0.77	0.91/0.95	0.97	0.93
	p=30	1/1/0.86	0.86	0.86	0.6/0.61	0.72/0.86	0.93	0.85
	p=60	1/1/0.84	-	0.84	0.53/0.54	0.57/0.85	0.9	0.83
	p=180	1/1/0.86	-	0.86	0.45/0.46	0.35/0.85	0.92	0.84
msnr	p=14	0.91/0.91/0.99	0.99	0.98	0.9/0.9	1/0.99	0.98	0.99
	p=30	0.91/0.91/0.99	0.97	0.99	0.84/0.85	0.9/0.99	1	0.96
	p=60	0.92/0.92/0.97	-	0.98	0.78/0.8	0.73/1	1	0.97
	p=180	0.91/0.91/0.96	-	0.96	0.76/0.77	0.49/1	1	0.97
lsnr	p=14	0.97/0.97/0.97	0.97	0.97	0.9/0.91	0.93/0.98	1	0.97
	p=30	1/1/0.92	0.91	0.92	0.74/0.76	0.71/0.93	0.98	0.92
	p=60	1/1/0.91	-	0.91	0.67/0.68	0.54/0.92	0.97	0.89
	p=180	1/1/0.9	-	0.9	0.57/0.58	0.32/0.9	0.95	0.88
		Sparsistency (number of extra variables)						
hsnr	p=14	6(0.4)/6(0.4)/6(0.6)	6(0.6)	6(0.6)	6(3.8)/6(4.5)	6(1.1)/6(0.8)	6(0.5)	6(0.7)
	p=30	6(0.1)/6(0.1)/6(0.6)	6(0.6)	6(0.6)	6(8.4)/6(8.7)	6(2.9)/6(1.3)	6(0.7)	6(0.9)
	p=60	6(0.1)/6(0.1)/6(0.5)	-	6(0.5)	6(13)/6(12.2)	6(6.8)/6(1.5)	6(1.1)	6(0.8)
	p=180	6(0)/6(0)/6(0.3)	-	6(0.3)	6(21.7)/6(19.4)	6(24.4)/6(1.5)	6(1.2)	6(0.6)
msnr	p=14	5.8(1.9)/5.8(1.8)/5.8(0.8)	5.8(0.8)	5.8(0.8)	6(3.8)/6(4.4)	5.9(1.2)/5.9(1.5)	5.8(1.1)	5.9(1)
	p=30	5.4(1)/5.4(0.9)/5.6(0.9)	5.6(0.9)	5.6(0.9)	6(8.5)/6(8.6)	5.9(3.4)/5.7(2.2)	5.7(1.9)	5.7(1.5)
	p=60	5.2(0.3)/5.2(0.3)/5.6(0.8)	-	5.6(0.8)	6(13)/6(12.1)	5.9(7.8)/5.7(2.6)	5.7(2.7)	5.6(1.3)
	p=180	4.7(0.1)/4.7(0.1)/5.2(0.6)	-	5.2(0.6)	5.9(21.7)/5.9(19)	5.9(26.9)/5.4(2.9)	5.4(4.5)	5.3(1.7)
lsnr	p=14	4.5(1)/4.5(1)/4.5(0.6)	4.5(0.6)	4.5(0.6)	5.3(3.4)/5.4(3.8)	4.8(1.3)/4.6(1)	4.5(0.7)	4.5(0.8)
	p=30	4.1(0.3)/4.1(0.3)/4.3(0.6)	4.3(0.6)	4.3(0.6)	5.1(7.2)/5.1(7.3)	4.8(3.6)/4.3(1.2)	4.3(1)	4.3(0.9)
	p=60	4.1(0.2)/4.1(0.2)/4.2(0.5)	-	4.2(0.5)	5(11.2)/4.9(10.1)	4.8(8)/4.2(1.2)	4.2(1.2)	4.2(1)
	p=180	4(0.1)/4(0.1)/4.1(0.4)	-	4.1(0.4)	4.7(18.7)/4.6(15.6)	4.8(27)/4.1(1)	4.1(1.4)	4.1(0.9)

Table S33: The performance of BOSS compared to other methods, Sparse-Ex4, ρ=0.5, n=200

		BOSS C_p-hdf/AICc-hdf/CV	BS CV	FS CV	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BOSS						
hsnr	p=14	34/33/25	22	26	29/28	17/25	25	32
	p=30	23/19/24	23	28	49/47	34/26	21	34
	p=60	18/17/21	-	24	56/54	52/21	17	31
	p=180	29/4/15	-	16	82/81	170/16	13	18
msnr	p=14	27/25/29	20	21	38/37	25/20	17	25
	p=30	24/19/34	24	30	74/72	52/26	16	56
	p=60	17/13/33	-	28	90/89	78/26	14	68
	p=180	52/14/37	-	29	137/137	263/29	17	103
lsnr	p=14	37/35/31	24	30	30/29	23/29	26	35
	p=30	37/35/32	23	34	35/35	30/31	23	40
	p=60	37/38/32	-	36	38/40	53/34	24	45
	p=180	49/31/26	-	34	33/35	120/31	20	40
		Relative efficiency						
hsnr	p=14	0.88/0.88/0.94	0.97	0.93	0.91/0.92	1/0.94	0.94	0.89
	p=30	0.97/1/0.96	0.97	0.93	0.8/0.81	0.89/0.95	0.99	0.89
	p=60	0.99/1/0.96	-	0.94	0.75/0.76	0.77/0.96	1	0.89
	p=180	0.81/1/0.91	-	0.9	0.57/0.58	0.39/0.9	0.92	0.89
msnr	p=14	0.92/0.94/0.91	0.97	0.96	0.84/0.85	0.93/0.97	1	0.93
	p=30	0.93/0.98/0.87	0.93	0.89	0.66/0.67	0.76/0.92	1	0.74
	p=60	0.97/1/0.85	-	0.88	0.6/0.6	0.64/0.9	1	0.67
	p=180	0.74/1/0.83	-	0.88	0.48/0.48	0.31/0.88	0.97	0.56
lsnr	p=14	0.9/0.91/0.94	0.99	0.95	0.95/0.95	1/0.95	0.98	0.91
	p=30	0.9/0.91/0.93	1	0.92	0.91/0.91	0.94/0.93	1	0.88
	p=60	0.91/0.9/0.94	-	0.91	0.9/0.89	0.81/0.93	1	0.86
	p=180	0.81/0.92/0.96	-	0.9	0.91/0.89	0.55/0.92	1	0.86
		Sparsistency (number of extra variables)						
hsnr	p=14	5.2(1.9)/5.1(1.4)/5.5(0.9)	5.5(0.9)	5.4(1)	5.9(4.6)/6(5.4)	5.7(1.3)/5.4(1.9)	5.4(1.4)	5.3(1.6)
	p=30	4.5(0.9)/4.4(0.2)/5(1)	5(1)	4.8(1.1)	5.7(10.4)/5.7(12)	5.4(3.6)/4.8(2.3)	4.8(2.1)	4.5(2)
	p=60	4.4(0.3)/4.2(0.1)/4.7(0.8)	-	4.6(0.8)	5.6(15)/5.7(17.8)	5.5(7.4)/4.7(2.3)	4.8(2.8)	4.4(2)
	p=180	4.3(11.4)/4(0)/4.2(0.5)	-	4.1(0.5)	5.1(20.2)/5.3(27.9)	5.2(35.3)/4.2(2.1)	4.2(3.5)	4.1(1.2)
msnr	p=14	4.7(1.6)/4.6(1.3)/4.4(0.6)	4.5(0.7)	4.5(0.7)	5.4(4.2)/5.5(5)	4.7(1.1)/4.6(1.2)	4.5(0.8)	4.6(1.2)
	p=30	4.2(1)/4.1(0.5)/4.2(0.7)	4.2(0.7)	4.2(1)	5(9.3)/5.1(10.7)	4.5(3.1)/4.2(1.4)	4.2(1.2)	4.4(3.4)
	p=60	4.1(0.4)/4(0.2)/4.1(0.6)	-	4.1(0.8)	4.8(13.6)/4.9(15.9)	4.5(6.8)/4.1(1.6)	4.1(1.8)	4.2(4.3)
	p=180	4.1(10.7)/4(0.1)/4(0.4)	-	4(0.6)	4.5(19.3)/4.6(26.2)	4.5(35.4)/4(1.7)	4.1(2.6)	4(5.6)
lsnr	p=14	4(2.7)/3.8(2.2)/3.3(0.8)	3.3(0.8)	3.3(1)	4.5(3.6)/4.7(4.1)	3.7(1.4)/3.5(1.7)	3.4(1.4)	3.6(2)
	p=30	2.7(2.3)/2.3(1)/2.7(1.3)	2.6(1)	2.6(1.5)	3.6(6.9)/3.6(7.5)	3.4(3.7)/2.9(3.3)	2.8(2.9)	2.9(4.1)
	p=60	1.9(1.1)/1.5(0.3)/2.3(1.2)	-	2(1.1)	2.9(8.5)/2.9(9.5)	3.4(7.8)/2.4(4)	2.5(3.9)	2.3(4.9)
	p=180	1.7(12.7)/1(0.2)/1.6(0.9)	-	1.3(0.8)	2.2(10.7)/2.1(13.3)	3.5(32.4)/1.9(6.4)	2.1(6.5)	1.8(7.6)

Table S34: The performance of BOSS compared to other methods, Sparse-Ex4, ρ=0.5, n=2000

		BOSS C_p-hdf/AICc-hdf/CV	BS CV	FS CV	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BOSS						
hsnr	p=14	12/11/25	19	18	48/47	19/16	13	18
	p=30	8/7/29	21	23	86/83	39/20	12	23
	p=60	7/7/28	-	23	125/120	73/22	14	24
	p=180	7/6/28	-	21	174/171	171/23	12	25
msnr	p=14	33/33/26	18	20	44/42	20/23	18	32
	p=30	40/39/33	21	31	71/69	43/35	26	54
	p=60	42/43/35	-	42	93/90	76/44	31	71
	p=180	41/42/33	-	46	105/102	153/48	37	69
lsnr	p=14	22/22/30	21	21	36/35	24/20	17	22
	p=30	14/13/30	22	23	61/59	55/20	15	24
	p=60	11/10/28	-	21	89/85	104/21	13	25
	p=180	8/8/27	-	20	125/121	242/19	10	29
		Relative efficiency						
hsnr	p=14	1/1/0.89	0.94	0.95	0.75/0.76	0.94/0.96	0.99	0.94
	p=30	1/1/0.83	0.89	0.87	0.58/0.59	0.77/0.89	0.96	0.88
	p=60	1/1/0.83	-	0.87	0.47/0.48	0.62/0.87	0.93	0.86
	p=180	1/1/0.83	-	0.88	0.39/0.39	0.39/0.87	0.95	0.85
msnr	p=14	0.88/0.88/0.93	1	0.98	0.82/0.83	0.99/0.96	1	0.89
	p=30	0.87/0.87/0.91	1	0.93	0.71/0.72	0.85/0.9	0.97	0.79
	p=60	0.92/0.92/0.98	-	0.92	0.68/0.69	0.75/0.91	1	0.77
	p=180	0.94/0.94/1	-	0.91	0.65/0.66	0.53/0.9	0.97	0.78
lsnr	p=14	0.96/0.96/0.9	0.97	0.96	0.86/0.87	0.94/0.98	1	0.96
	p=30	0.99/1/0.87	0.93	0.92	0.7/0.71	0.73/0.95	0.99	0.91
	p=60	1/1/0.86	-	0.91	0.58/0.59	0.54/0.91	0.98	0.88
	p=180	0.99/1/0.85	-	0.9	0.48/0.49	0.32/0.91	0.98	0.83
		Sparsistency (number of extra variables)						
hsnr	p=14	6(0.6)/6(0.5)/6(0.6)	6(0.7)	6(0.6)	6(4.7)/6(5.5)	6(1)/6(1)	6(0.6)	6(0.7)
	p=30	6(0.2)/6(0.2)/6(0.6)	6(0.6)	6(0.6)	6(10.9)/6(11.3)	6(2.7)/6(1.4)	6(0.8)	6(0.7)
	p=60	6(0.1)/6(0.1)/6(0.5)	-	6(0.5)	6(18)/6(17.7)	6(6.4)/6(1.5)	6(1.1)	6(0.7)
	p=180	6(0.1)/6(0.1)/6(0.3)	-	6(0.3)	6(32.2)/6(29.8)	6(22.3)/6(1.3)	6(1.3)	6(0.5)
msnr	p=14	5.9(2.1)/5.9(2)/6(0.7)	5.9(0.7)	5.9(0.9)	6(4.8)/6(5.5)	6(1.2)/5.9(1.9)	5.9(1.1)	5.9(1.6)
	p=30	5.6(1.1)/5.6(1)/5.9(0.8)	5.9(0.7)	5.8(1)	6(10.9)/6(11.3)	6(3.3)/5.8(2.8)	5.8(2)	5.7(2.3)
	p=60	5.4(0.4)/5.3(0.3)/5.8(0.8)	-	5.6(1.1)	6(18)/6(17.6)	6(7.7)/5.6(3.7)	5.7(3.1)	5.4(2.6)
	p=180	4.8(0.2)/4.8(0.2)/5.5(1)	-	5.1(0.8)	5.9(31.7)/5.9(29)	6(26.5)/5.1(4.5)	5.2(4.7)	4.8(2.1)
lsnr	p=14	4.6(1.4)/4.6(1.3)/4.6(0.7)	4.6(0.7)	4.6(0.8)	5.5(4.4)/5.5(5)	4.9(1.3)/4.6(1.2)	4.5(0.8)	4.5(0.9)
	p=30	4.2(0.5)/4.2(0.4)/4.3(0.7)	4.3(0.6)	4.2(0.7)	5.2(9.6)/5.3(9.9)	4.8(3.6)/4.3(1.1)	4.3(1)	4.2(1)
	p=60	4.1(0.2)/4.1(0.2)/4.2(0.5)	-	4.1(0.5)	5(15.5)/5(14.7)	4.8(8.2)/4.1(1.2)	4.1(1.2)	4.1(1)
	p=180	4(0.1)/4(0.1)/4(0.4)	-	4(0.4)	4.6(26.1)/4.6(23.6)	4.8(27.9)/4(1)	4.1(1.3)	4(1)

Table S35: The performance of BOSS compared to other methods, Sparse-Ex4, ρ=0.9, n=200

		BOSS C_p-hdf/AICc-hdf/CV	BS CV	FS CV	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BOSS						
hsnr	p=14	34/33/29	24	46	45/44	33/43	40	54
	p=30	25/21/33	21	56	73/71	40/36	28	67
	p=60	23/22/37	-	57	95/91	50/31	22	78
	p=180	27/7/34	-	68	123/110	91/2	-10	81
msnr	p=14	29/27/28	15	33	41/39	23/28	16	41
	p=30	25/19/30	-4	52	61/60	29/23	3	59
	p=60	22/14/24	-	83	90/86	57/38	16	93
	p=180	31/17/26	-	66	78/66	78/26	6	72
lsnr	p=14	41/39/31	27	54	51/50	46/47	38	56
	p=30	39/32/27	18	78	71/72	58/66	44	78
	p=60	29/28/23	-	81	80/80	79/79	57	86
	p=180	37/17/18	-	36	33/33	110/37	35	37
		Relative efficiency						
hsnr	p=14	0.93/0.94/0.96	1	0.85	0.86/0.86	0.94/0.87	0.89	0.81
	p=30	0.97/1/0.91	1	0.78	0.7/0.71	0.87/0.9	0.95	0.73
	p=60	0.99/1/0.89	-	0.77	0.62/0.64	0.81/0.93	0.99	0.68
	p=180	0.71/0.84/0.68	-	0.54	0.41/0.43	0.47/0.88	1	0.5
msnr	p=14	0.89/0.91/0.9	1	0.87	0.82/0.83	0.93/0.9	0.99	0.82
	p=30	0.77/0.81/0.74	1	0.63	0.6/0.6	0.75/0.78	0.93	0.6
	p=60	0.93/1/0.92	-	0.62	0.6/0.61	0.73/0.83	0.99	0.59
	p=180	0.81/0.9/0.84	-	0.64	0.59/0.64	0.59/0.84	1	0.61
lsnr	p=14	0.9/0.91/0.97	1	0.83	0.84/0.84	0.87/0.86	0.92	0.81
	p=30	0.85/0.89/0.93	1	0.66	0.69/0.69	0.75/0.71	0.82	0.66
	p=60	0.95/0.97/1	-	0.68	0.69/0.68	0.69/0.69	0.78	0.66
	p=180	0.86/1/0.99	-	0.86	0.88/0.88	0.56/0.86	0.87	0.85
		Sparsistency (number of extra variables)						
hsnr	p=14	5.8(2.9)/5.7(2.5)/5.6(1.6)	5.6(0.8)	5.3(2)	6(6.5)/6(7.2)	5.5(2.3)/5.4(3.4)	5.3(2.4)	5.5(3.9)
	p=30	5.2(3.7)/5.1(2.8)/5.3(3.8)	5(1)	4.8(4.1)	5.8(17.8)/5.8(19.5)	4.9(4.5)/4.5(3.4)	4.4(2.7)	5(9.2)
	p=60	5(4.2)/4.8(3)/4.9(4.7)	-	4.5(4.3)	5.6(30.2)/5.7(35.1)	4.9(8.8)/4.3(3.6)	4.3(3.3)	4.5(12.5)
	p=180	4.4(13.2)/4.2(2.4)/4.3(4.3)	-	4.3(8)	4.6(44.2)/4.8(64.9)	4.4(29.2)/4.1(4)	4.1(3.1)	4.2(20.4)
msnr	p=14	5(2.5)/4.9(2.2)/4.7(1.4)	4.6(0.7)	4.9(2.4)	5.6(6.3)/5.7(7)	4.6(1.8)/4.9(3.3)	4.6(2.4)	5.1(4.4)
	p=30	4.5(4.5)/4.4(3.6)/4.5(3.7)	4.2(0.7)	4.6(6.5)	5.2(16.9)/5.3(18.5)	4.4(4.7)/4.5(7)	4.4(5.8)	4.9(12.7)
	p=60	4.3(5.9)/4.2(5.1)/4.3(5.7)	-	3.7(6.3)	4.8(28.3)/5(32.8)	4.4(10.4)/4.2(11.8)	4.3(11.7)	4.2(20)
	p=180	4.1(24.5)/3.6(7.6)/3.7(9.2)	-	3(8.8)	3.8(37.7)/4.3(59.6)	4.3(37.5)/3.6(20.4)	3.8(22.8)	3.3(28.7)
lsnr	p=14	4.2(3)/4(2.7)/3.8(1.9)	3.3(0.7)	3.9(2.9)	4.9(5.9)/5.1(6.5)	3.8(2.6)/4.1(4.1)	3.7(3.2)	4.5(5.1)
	p=30	3.1(5.5)/2.7(3.9)/3.2(5)	2.7(0.9)	2.2(4.4)	3(10.1)/3(11)	3.3(7.6)/2.8(8)	2.9(8)	2.6(8.7)
	p=60	2.2(5.3)/1.9(3.9)/2.6(6.8)	-	0.9(2)	1.2(7.4)/1.2(8.4)	2.8(12.8)/1.3(6.6)	2(10)	1.1(6.6)
	p=180	1.4(18.2)/0.7(1.7)/1.1(5.5)	-	0.2(0.6)	0.3(4.8)/0.3(5.4)	2.3(40.2)/0.3(4.4)	0.6(9)	0.3(3.9)

Table S36: The performance of BOSS compared to other methods, Sparse-Ex4, ρ=0.9, n=2000

		BOSS C_p-hdf/AICc-hdf/CV	BS CV	FS CV	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BOSS						
hsnr	p=14	33/33/27	16	20	53/51	17/20	15	31
	p=30	33/32/33	16	33	108/106	26/22	12	55
	p=60	27/26/30	-	48	157/154	37/22	9	86
	p=180	15/15/25	-	90	226/223	68/33	10	154
msnr	p=14	28/28/22	18	45	53/51	26/39	30	57
	p=30	22/22/26	23	89	108/105	56/75	55	112
	p=60	21/21/27	-	114	166/162	100/107	92	128
	p=180	25/25/30	-	107	253/250	190/110	105	110
lsnr	p=14	33/33/32	20	29	45/44	24/28	19	40
	p=30	28/27/34	16	40	85/82	44/30	12	63
	p=60	19/19/30	-	59	122/119	75/32	9	95
	p=180	15/14/27	-	104	179/176	167/48	20	164
		Relative efficiency						
hsnr	p=14	0.86/0.86/0.9	0.99	0.96	0.75/0.76	0.98/0.95	1	0.87
	p=30	0.84/0.84/0.84	0.96	0.84	0.54/0.54	0.89/0.92	1	0.72
	p=60	0.86/0.86/0.84	-	0.74	0.42/0.43	0.8/0.89	1	0.58
	p=180	0.95/0.96/0.88	-	0.58	0.34/0.34	0.65/0.82	1	0.43
msnr	p=14	0.92/0.92/0.97	1	0.81	0.77/0.78	0.94/0.85	0.91	0.75
	p=30	0.99/1/0.96	0.99	0.64	0.58/0.59	0.78/0.7	0.78	0.58
	p=60	1/1/0.95	-	0.56	0.45/0.46	0.6/0.58	0.63	0.53
	p=180	1/1/0.96	-	0.6	0.35/0.36	0.43/0.6	0.61	0.6
lsnr	p=14	0.9/0.9/0.9	1	0.92	0.82/0.83	0.96/0.93	1	0.85
	p=30	0.87/0.88/0.83	0.97	0.8	0.61/0.61	0.78/0.86	1	0.69
	p=60	0.92/0.92/0.84	-	0.69	0.49/0.5	0.62/0.83	1	0.56
	p=180	1/1/0.9	-	0.56	0.41/0.41	0.43/0.77	0.95	0.43
		Sparsistency (number of extra variables)						
hsnr	p=14	6(1.4)/6(1.4)/6(0.6)	6(0.7)	6(0.9)	6(6.6)/6(7.2)	6(1.2)/6(1.8)	6(0.9)	6(1.8)
	p=30	6(0.5)/6(0.4)/6(0.7)	6(0.6)	6(1.3)	6(17.8)/6(18.9)	6(2.6)/6(2.5)	6(1.6)	6(3)
	p=60	6(0.4)/6(0.4)/6(0.8)	-	6(2)	6(34.2)/6(35.3)	6(5.8)/6(3.9)	6(3.1)	6(5.1)
	p=180	6(1.4)/6(1.4)/6(1.7)	-	5.9(3.8)	6(72.7)/6(73.8)	6(17.2)/6(9.4)	6(8.7)	5.9(11.5)
msnr	p=14	6(2)/6(2)/6(1.3)	6(0.7)	5.9(2.4)	6(6.6)/6(7.2)	6(2.4)/5.9(4.1)	5.9(2.9)	5.8(4.2)
	p=30	5.9(2.4)/5.9(2.4)/5.9(2.6)	5.9(0.7)	5.4(3.4)	6(17.8)/6(18.8)	5.9(5.9)/5.6(7.2)	5.6(6.2)	5.2(5.9)
	p=60	5.8(4.8)/5.8(4.8)/5.9(5.6)	-	4.7(1.9)	6(34)/6(35.1)	5.8(12.4)/4.9(7.7)	5(8.3)	4.4(3.1)
	p=180	5.5(12.9)/5.5(12.7)/5.7(16.2)	-	4.2(0.6)	5.7(68.6)/5.7(69.3)	5.4(31.9)/4.2(3.4)	4.3(5)	4.1(0.6)
lsnr	p=14	5(2.5)/5(2.4)/4.7(1.1)	4.6(0.7)	4.6(1.5)	5.7(6.4)/5.7(7)	4.6(1.5)/4.7(2.4)	4.5(1.5)	4.7(2.8)
	p=30	4.5(2.1)/4.4(1.9)/4.4(1.7)	4.3(0.6)	4.3(2.2)	5.4(16.6)/5.4(17.6)	4.5(3.7)/4.3(3.5)	4.2(2.5)	4.5(5.3)
	p=60	4.2(1.9)/4.2(1.8)/4.3(2.3)	-	4.2(3.2)	5.1(30.6)/5.1(31.3)	4.3(8.1)/4.2(5.8)	4.2(5)	4.4(9.8)
	p=180	4.1(3.7)/4.1(3.6)/4.1(4.4)	-	3.7(3.7)	4.6(60.3)/4.6(60.5)	4.3(26.8)/4.1(14.8)	4.2(14.2)	3.7(15.1)

Table S37: The performance of BOSS compared to other methods, Dense, ρ=0, n=200

		BOSS C_p-hdf/AICc-hdf/CV	BS CV	FS CV	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BOSS						
hsnr	p=14	0/0/0	0	0	0/0	0/1	0	75
	p=30	2/2/7	8	7	1/1	8/2	4	5
	p=60	9/11/11	-	11	0/-2	-1/-1	1	3
	p=180	18/13/11	-	12	18/12	5/5	3	13
msnr	p=14	0/0/6	6	6	1/1	7/1	3	21
	p=30	4/5/10	10	10	0/-1	8/2	4	4
	p=60	11/12/12	-	12	-3/-5	-1/-2	-1	0
	p=180	19/14/13	-	12	3/0	13/1	2	6
lsnr	p=14	7/9/20	19	20	3/3	17/8	10	13
	p=30	15/19/16	16	16	-8/-8	5/-1	-3	-1
	p=60	16/14/14	-	14	-9/-8	11/-4	-4	-3
	p=180	22/7/8	-	9	-5/-3	65/0	-1	2
		Relative efficiency						
hsnr	p=14	1/1/1	1	1	1/1	1/0.99	1	0.57
	p=30	0.99/0.99/0.95	0.94	0.94	1/1	0.93/0.99	0.97	0.97
	p=60	0.9/0.88/0.89	-	0.89	0.98/1	1/0.99	0.98	0.96
	p=180	0.88/0.91/0.93	-	0.93	0.88/0.93	0.99/0.99	1	0.92
msnr	p=14	1/1/0.95	0.95	0.95	1/1	0.94/0.99	0.97	0.83
	p=30	0.96/0.95/0.91	0.9	0.91	0.99/1	0.92/0.98	0.96	0.95
	p=60	0.86/0.85/0.85	-	0.85	0.98/1	0.96/0.97	0.96	0.96
	p=180	0.84/0.88/0.89	-	0.89	0.97/1	0.88/0.98	0.98	0.94
lsnr	p=14	0.96/0.94/0.86	0.86	0.85	0.99/1	0.88/0.95	0.93	0.91
	p=30	0.8/0.77/0.79	0.79	0.79	1/1	0.88/0.93	0.94	0.93
	p=60	0.79/0.8/0.8	-	0.8	1/1	0.83/0.95	0.96	0.94
	p=180	0.78/0.89/0.88	-	0.88	1/0.98	0.58/0.95	0.96	0.94
		Sparsistency (number of extra variables)						
hsnr	p=14	14/14/14	14	14	14/14	14/14	14	13
	p=30	29.2/29/26	25.7	25.9	28.6/29.2	25.3/28.6	27.2	26.3
	p=60	35.8/24.3/28	-	27.6	40.5/44.9	29.8/38.4	36.8	32.4
	p=180	36.9/17/19.3	-	19.2	47.5/67.6	38.7/36.4	32.4	36.3
msnr	p=14	14/14/13.4	13.4	13.4	13.9/13.9	13.4/13.9	13.6	12.7
	p=30	26.9/26/21.1	20.7	20.9	25.1/26.3	18.8/24.9	23.7	22.1
	p=60	26/14.9/18.4	-	18.3	32.2/36.4	22.7/29.2	29.6	24.2
	p=180	30.7/7.9/9.9	-	9.9	33.8/49.1	40.6/30.1	27.9	24.3
lsnr	p=14	12.2/11.8/8.3	8.4	8.2	10.6/11.2	7.6/10.4	10	8.6
	p=30	13/7.8/8	7.5	7.8	14/15.1	10.3/12.3	12.7	10.3
	p=60	5.4/1/4.2	-	4.1	15.2/16.6	15.3/13.7	14.6	11.1
	p=180	16.6/0.3/1.4	-	1.5	12.4/15.8	41.6/12.2	13.5	10.7

Table S38: The performance of BOSS compared to other methods, Dense, ρ=0, n=2000

		BOSS C_p-hdf/AICc-hdf/CV	BS CV	FS CV	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BOSS						
hsnr	p=14	0/0/0	0	0	2/2	1/4	0	295
	p=30	0/0/1	1	1	1/1	3/6	1	16
	p=60	5/5/7	-	7	1/0	0/0	3	5
	p=180	7/7/9	-	9	15/14	3/5	2	9
msnr	p=14	0/0/0	0	0	0/0	0/2	0	105
	p=30	1/1/5	5	5	1/1	4/2	3	7
	p=60	8/8/9	-	9	1/0	-1/0	3	4
	p=180	8/8/10	-	10	13/12	3/4	3	7
lsnr	p=14	0/0/4	4	4	0/0	3/1	3	27
	p=30	3/3/10	10	9	1/0	7/2	5	7
	p=60	13/13/12	-	12	0/-1	0/1	3	4
	p=180	9/9/12	-	13	9/8	14/5	4	8
		Relative efficiency						
hsnr	p=14	1/1/1	1	1	0.98/0.98	0.99/0.96	1	0.25
	p=30	1/1/0.99	0.99	0.99	0.99/0.99	0.97/0.94	0.99	0.86
	p=60	0.95/0.95/0.93	-	0.93	0.99/0.99	1/1	0.97	0.95
	p=180	0.95/0.96/0.94	-	0.94	0.89/0.9	0.99/0.98	1	0.94
msnr	p=14	1/1/1	1	1	1/1	1/0.99	1	0.49
	p=30	1/1/0.96	0.96	0.96	1/1	0.97/0.99	0.97	0.94
	p=60	0.92/0.92/0.91	-	0.91	0.98/0.99	1/0.99	0.96	0.95
	p=180	0.95/0.95/0.93	-	0.93	0.91/0.92	1/0.99	1	0.96
lsnr	p=14	1/1/0.96	0.96	0.96	1/1	0.98/0.99	0.98	0.79
	p=30	0.97/0.97/0.91	0.91	0.92	1/1	0.94/0.98	0.96	0.94
	p=60	0.88/0.88/0.88	-	0.88	0.99/1	0.99/0.99	0.96	0.95
	p=180	0.95/0.96/0.93	-	0.93	0.96/0.97	0.91/1	1	0.97
		Sparsistency (number of extra variables)						
hsnr	p=14	14/14/14	14	14	14/14	14/14	14	13
	p=30	30/30/29.8	29.8	29.8	30/30	30/30	29.9	28.8
	p=60	50.1/49.6/39.8	-	39.7	53.1/54.3	46.4/50.2	44.8	41.5
	p=180	32.5/31.6/32.3	-	32.1	88.1/88.8	62.2/60.9	46.4	37.9
msnr	p=14	14/14/14	14	14	14/14	14/14	14	13
	p=30	29.8/29.8/28.1	28.2	28.1	29.6/29.8	28.9/29.7	28.7	27.8
	p=60	42/41.1/31.3	-	31.1	48.1/49.1	37.1/43.5	37.5	33.4
	p=180	23.9/23.2/24.1	-	24.1	75.1/74.9	51.3/44.6	38.9	27.9
lsnr	p=14	14/14/13.7	13.7	13.7	14/14	13.8/14	13.8	12.7
	p=30	27.8/27.8/22.1	21.9	22.1	26.8/27.5	21/26.2	24.6	22.4
	p=60	30/28.5/19.9	-	19.5	38.5/38.9	25.4/31.4	29.1	24.2
	p=180	14.1/13.5/14.1	-	14	55.3/53.9	42/31.8	29.7	22.2

Table S39: The performance of BOSS compared to other methods, Dense, ρ=0.5, n=200

		BOSS C_p-hdf/AICc-hdf/CV	BS CV	FS CV	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BOSS						
hsnr	p=14	0/0/4	0	0	0/0	1/2	0	113
	p=30	1/1/8	9	8	2/1	6/2	5	7
	p=60	13/13/12	-	12	10/7	5/5	7	9
	p=180	37/14/13	-	16	47/27	10/13	8	28
msnr	p=14	0/0/5	5	4	0/0	2/1	2	32
	p=30	4/5/10	11	10	4/3	8/4	8	6
	p=60	16/15/14	-	14	11/6	4/6	7	8
	p=180	43/14/14	-	17	37/21	18/18	11	26
lsnr	p=14	5/7/17	19	17	9/7	16/8	14	15
	p=30	18/18/17	16	18	10/7	10/10	11	12
	p=60	17/13/13	-	14	6/5	19/8	8	9
	p=180	47/4/8	-	9	2/4	79/6	6	8
		Relative efficiency						
hsnr	p=14	1/1/0.96	1	1	1/1	0.99/0.98	1	0.47
	p=30	0.98/0.98/0.93	0.91	0.93	0.98/0.98	0.94/0.98	0.94	0.93
	p=60	0.93/0.93/0.94	-	0.93	0.96/0.99	1/1	0.98	0.97
	p=180	0.79/0.95/0.96	-	0.93	0.73/0.85	0.98/0.95	1	0.84
msnr	p=14	1/1/0.95	0.95	0.96	1/1	0.98/0.99	0.98	0.76
	p=30	0.98/0.98/0.93	0.92	0.93	0.99/1	0.95/0.98	0.95	0.97
	p=60	0.89/0.91/0.92	-	0.91	0.94/0.98	1/0.98	0.97	0.96
	p=180	0.77/0.97/0.97	-	0.95	0.81/0.91	0.94/0.94	1	0.88
lsnr	p=14	0.97/0.96/0.87	0.86	0.87	0.93/0.96	0.88/0.94	0.9	0.89
	p=30	0.91/0.91/0.92	0.92	0.91	0.97/1	0.98/0.97	0.96	0.96
	p=60	0.9/0.93/0.93	-	0.93	0.99/1	0.88/0.98	0.97	0.97
	p=180	0.69/0.97/0.94	-	0.94	1/0.98	0.57/0.96	0.96	0.95
		Sparsistency (number of extra variables)						
hsnr	p=14	14/14/14	14	14	14/14	14/14	14	13
	p=30	29.7/29.6/26.1	25.1	26	29.1/29.7	27/29.2	27	27.5
	p=60	42/29.1/29.4	-	28.6	44.7/50.7	33.7/43.1	35	42.8
	p=180	43.6/17/20.2	-	19.6	52.2/87.1	40.3/42.5	32.4	62.4
msnr	p=14	14/14/13.7	13.6	13.7	14/14	13.8/14	13.8	12.9
	p=30	28.3/27.8/21.1	18.8	20.9	26.7/28.2	20.3/26.2	22.6	25.1
	p=60	34.4/17.9/19.9	-	19.6	34.7/43.5	24.7/33.7	27.1	36.5
	p=180	36.7/8.3/11	-	9.7	24.5/61.4	39.6/28.9	22.1	44.9
lsnr	p=14	13.1/12.7/9.4	8.7	9.3	10.8/12	8.7/11.4	10.3	10.1
	p=30	13.4/5.7/7.5	6.6	7.1	5.3/12.9	11.8/11.8	10.3	10.2
	p=60	4.8/0.6/3.5	-	3	3.7/11.2	15.7/9.4	8.9	7.6
	p=180	19/0.2/1.1	-	0.9	2.7/8.6	38.9/7.2	6.9	5.8

Table S40: The performance of BOSS compared to other methods, Dense, ρ=0.5, n=2000

		BOSS C_p-hdf/AICc-hdf/CV	BS CV	FS CV	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BOSS						
hsnr	p=14	0/0/0	0	0	2/2	2/5	0	452
	p=30	0/0/3	0	0	1/1	9/10	0	26
	p=60	6/6/7	-	7	4/3	4/5	5	7
	p=180	8/8/9	-	12	37/36	20/20	8	38
msnr	p=14	0/0/2	0	0	0/0	1/4	0	179
	p=30	0/0/5	6	5	1/0	3/5	3	9
	p=60	10/10/10	-	10	7/6	4/5	7	10
	p=180	9/9/11	-	14	36/34	16/18	9	37
lsnr	p=14	0/0/5	2	2	0/0	1/1	1	45
	p=30	2/2/9	10	10	2/1	7/2	7	6
	p=60	17/16/13	-	14	11/10	6/8	8	13
	p=180	11/11/13	-	16	32/30	19/18	10	34
		Relative efficiency						
hsnr	p=14	1/1/1	1	1	0.98/0.98	0.98/0.95	1	0.18
	p=30	0.98/0.98/0.96	0.98	0.98	0.97/0.97	0.9/0.89	0.98	0.78
	p=60	0.98/0.98/0.96	-	0.96	1/1	0.99/0.98	0.98	0.97
	p=180	1/1/0.99	-	0.96	0.79/0.79	0.9/0.89	1	0.78
msnr	p=14	1/1/0.98	1	1	1/1	0.99/0.96	1	0.36
	p=30	0.98/0.98/0.94	0.93	0.94	0.98/0.98	0.95/0.94	0.96	0.91
	p=60	0.95/0.95/0.95	-	0.95	0.98/0.98	1/1	0.98	0.95
	p=180	0.99/1/0.98	-	0.95	0.8/0.81	0.93/0.92	1	0.79
lsnr	p=14	1/1/0.95	0.98	0.98	1/1	0.99/0.99	0.99	0.69
	p=30	0.98/0.98/0.91	0.9	0.91	0.98/0.98	0.93/0.97	0.93	0.94
	p=60	0.91/0.91/0.93	-	0.93	0.95/0.96	1/0.98	0.98	0.93
	p=180	0.99/1/0.97	-	0.95	0.83/0.85	0.93/0.93	1	0.83
		Sparsistency (number of extra variables)						
hsnr	p=14	14/14/14	14	14	14/14	14/14	14	13
	p=30	30/30/30	30	30	30/30	30/30	30	28.8
	p=60	53.6/53.3/40.3	-	40	55.4/56.9	48.4/49.4	43.8	48.4
	p=180	36/34.5/35.1	-	32.6	106.5/113.5	76.5/77.1	43	63.7
msnr	p=14	14/14/14	14	14	14/14	14/14	14	13
	p=30	30/30/28.6	28.3	28.5	29.8/29.9	29.4/29.7	29.1	28.2
	p=60	47.5/46.6/31.6	-	31.2	51.5/53.4	42.1/48.1	36.3	43
	p=180	27.3/26.2/27.1	-	24.1	90.5/98.9	60/54.4	35.2	53.6
lsnr	p=14	14/14/13.9	13.9	13.9	14/14	14/14	13.9	12.8
	p=30	29.1/29/22.7	21.2	22.1	28/28.9	23/27.6	24.1	25.5
	p=60	36.1/34.7/21.1	-	19.8	42.2/45.5	28.3/34.9	26.5	34.9
	p=180	17/16/17	-	14.3	61.8/72	43.7/33.1	25.3	43.6

Table S41: The performance of BOSS compared to other methods, Dense, ρ=0.9, n=200

		BOSS C_p-hdf/AICc-hdf/CV	BS CV	FS CV	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BOSS						
hsnr	p=14	0/0/5	0	0	0/0	6/7	1	64
	p=30	2/2/9	10	8	2/2	14/15	8	9
	p=60	15/19/14	-	12	12/11	12/13	14	14
	p=180	46/15/12	-	12	71/43	23/24	20	46
msnr	p=14	1/1/7	8	6	2/1	8/8	5	24
	p=30	4/5/12	12	10	4/3	8/5	9	7
	p=60	20/23/16	-	13	30/17	12/15	15	19
	p=180	58/26/15	-	17	61/46	31/37	26	49
lsnr	p=14	11/13/18	19	18	11/10	13/12	18	16
	p=30	19/15/14	15	13	10/11	19/12	12	11
	p=60	14/11/12	-	12	9/12	34/13	12	12
	p=180	59/12/10	-	10	15/17	106/12	9	11
		Relative efficiency						
hsnr	p=14	0.99/0.99/0.95	0.99	0.99	0.99/0.99	0.94/0.93	0.99	0.61
	p=30	0.98/0.98/0.91	0.9	0.92	0.98/0.98	0.87/0.87	0.93	0.92
	p=60	0.87/0.84/0.87	-	0.89	0.89/0.9	0.89/0.88	0.88	0.88
	p=180	0.77/0.97/1	-	1	0.65/0.78	0.91/0.9	0.93	0.77
msnr	p=14	0.99/0.99/0.94	0.93	0.94	0.98/0.99	0.93/0.93	0.95	0.8
	p=30	0.99/0.98/0.92	0.92	0.93	0.99/1	0.95/0.98	0.95	0.97
	p=60	0.93/0.92/0.97	-	0.99	0.86/0.96	1/0.97	0.98	0.94
	p=180	0.73/0.91/1	-	0.98	0.71/0.79	0.88/0.84	0.91	0.77
lsnr	p=14	0.93/0.91/0.87	0.87	0.88	0.94/0.94	0.91/0.93	0.88	0.9
	p=30	0.92/0.96/0.97	0.96	0.97	1/1	0.92/0.98	0.98	0.99
	p=60	0.95/0.98/0.97	-	0.97	1/0.97	0.81/0.97	0.97	0.97
	p=180	0.69/0.97/0.99	-	1	0.95/0.93	0.53/0.98	1	0.98
		Sparsistency (number of extra variables)						
hsnr	p=14	14/14/14	14	14	14/14	14/14	14	12.7
	p=30	29.5/29.3/25.2	23	24.6	28.9/29.5	24.4/24.8	26.2	26.7
	p=60	44.3/26.1/27.4	-	23.6	46.5/49.8	31.6/34.4	32.9	42.1
	p=180	46.7/15.6/21.3	-	17.2	54.4/93.4	44.8/46.9	37.7	71.4
msnr	p=14	14/14/13.5	13.2	13.4	13.9/14	13.4/13.7	13.5	12.5
	p=30	28.6/28/20.2	16.6	19	26.4/27.9	19.3/24.4	22.3	25
	p=60	33/13/18.4	-	14.9	30.3/42.2	24.7/32.3	25.8	36.5
	p=180	41.6/5.8/14.1	-	9.3	4.8/42.5	46.7/30.9	26	30.2
lsnr	p=14	10.6/9.8/7.1	6.4	6.6	8.9/9.3	7.5/8.6	7	7.5
	p=30	8.6/4.1/5.2	4.3	4.5	8.4/9.2	10.3/7.6	5.9	6.7
	p=60	3.7/1.4/4	-	2.7	2.7/8.1	14.5/6.2	5	5
	p=180	18/1/2.1	-	1.6	3.9/6.8	37.7/4	3.3	3

Table S42: The performance of BOSS compared to other methods, Dense, ρ=0.9, n=2000

		BOSS C_p-hdf/AICc-hdf/CV	BS CV	FS CV	lasso AICc/CV	gamma lasso AICc/CV	SparseNet CV	rlasso CV
		% worse than the best possible BOSS						
hsnr	p=14	0/0/0	0	0	3/4	27/27	2	380
	p=30	0/0/3	1	1	2/2	106/106	1	24
	p=60	7/8/9	-	6	6/6	74/74	8	8
	p=180	10/10/10	-	9	39/38	56/56	17	39
msnr	p=14	0/0/3	0	0	1/1	8/8	1	141
	p=30	0/0/5	7	5	1/1	25/25	3	8
	p=60	12/12/10	-	8	9/8	11/11	9	11
	p=180	10/11/12	-	9	41/39	22/23	16	41
lsnr	p=14	1/1/7	5	4	1/1	5/8	3	34
	p=30	3/3/11	11	9	3/2	7/6	8	7
	p=60	19/18/14	-	10	14/13	8/11	12	16
	p=180	12/12/13	-	11	40/38	27/26	18	39
		Relative efficiency						
hsnr	p=14	1/1/1	1	1	0.97/0.97	0.79/0.79	0.98	0.21
	p=30	1/1/0.97	0.99	0.99	0.98/0.98	0.49/0.49	0.99	0.8
	p=60	0.98/0.98/0.97	-	1	1/1	0.61/0.61	0.97	0.97
	p=180	0.99/0.99/0.99	-	1	0.78/0.79	0.7/0.7	0.93	0.78
msnr	p=14	1/1/0.97	1	1	0.99/0.99	0.92/0.92	0.99	0.42
	p=30	1/1/0.96	0.94	0.95	1/1	0.8/0.8	0.97	0.93
	p=60	0.96/0.96/0.98	-	1	0.99/0.99	0.97/0.97	0.99	0.97
	p=180	0.98/0.98/0.98	-	1	0.77/0.78	0.89/0.88	0.94	0.77
lsnr	p=14	1/1/0.94	0.95	0.97	1/1	0.96/0.93	0.98	0.75
	p=30	0.99/0.99/0.92	0.92	0.94	1/1	0.96/0.96	0.94	0.96
	p=60	0.91/0.92/0.95	-	0.99	0.95/0.96	1/0.98	0.97	0.94
	p=180	1/0.99/0.99	-	1	0.8/0.81	0.88/0.88	0.94	0.8
		Sparsistency (number of extra variables)						
hsnr	p=14	14/14/14	14	14	14/14	14/14	14	12.8
	p=30	30/30/29.9	29.9	29.9	30/30	26.1/26.1	30	28.4
	p=60	53.2/52.7/39.2	-	36.3	55.5/57.2	29.4/29.4	46	49.5
	p=180	37/35/38.6	-	30.2	109.6/118.6	43.1/43.1	52.4	75.6
msnr	p=14	14/14/14	14	14	14/14	14/14	14	12.9
	p=30	29.9/29.9/28.2	27.3	27.9	29.7/29.9	26.3/26.3	28.8	27.9
	p=60	47.5/46.6/31.3	-	27.4	51.1/53.4	35.5/35.6	37.2	44.7
	p=180	30.8/28.4/32	-	23	95.1/104.1	62.7/59	44.6	72
lsnr	p=14	14/14/13.8	13.6	13.7	14/14	13.8/13.9	13.8	12.7
	p=30	28.8/28.8/21.2	18.5	20.2	27.6/28.4	21.7/24.8	23.5	25.1
	p=60	35.3/33.6/20.6	-	16.3	41/43.9	28.5/33.1	26.4	36.2
	p=180	18.5/16.6/21.4	-	11.8	65.3/76.1	49.6/41.2	32.3	61.8

References

Csorgo, S., E. Haeusler, and D. M. Mason (1991). The asymptotic distribution of extreme sums. *The Annals of Probability 19(2)*, 783–811.

Embrechts, P., C. Klüppelberg, and T. Mikosch (2013). *Modelling Extremal Events: for Insurance and Finance*, Volume 33. Berlin: Springer Science & Business Media.

Fung, T. and E. Seneta (2017). Quantile function expansion using regularly varying functions. *Methodology and Computing in Applied Probability 20(4)*, 1091–1103.

Taddy, M. (2017). One-step estimator paths for concave regularization. *Journal of Computational and Graphical Statistics 26*(3), 525–536.

Zou, H., T. Hastie, and R. Tibshirani (2007). On the "degrees of freedom" of the lasso. *The Annals of Statistics 35*(5), 2173–2192.