The Annals of Statistics
2008, Vol. 36, No. 4, 1595–1618
DOI: 10.1214/009053607000000578

"PRECONDITIONING" FOR FEATURE SELECTION AND REGRESSION IN HIGH-DIMENSIONAL PROBLEMS[1]

BY DEBASHIS PAUL, ERIC BAIR, TREVOR HASTIE[1] AND ROBERT TIBSHIRANI[2]

University of California, Davis, Stanford University, Stanford University and Stanford University

We consider regression problems where the number of predictors greatly exceeds the number of observations. We propose a method for variable selection that first estimates the regression function, yielding a "preconditioned" response variable. The primary method used for this initial regression is supervised principal components. Then we apply a standard procedure such as forward stepwise selection or the LASSO to the preconditioned response variable. In a number of simulated and real data examples, this two-step procedure outperforms forward stepwise selection or the usual LASSO (applied directly to the raw outcome). We also show that under a certain Gaussian latent variable model, application of the LASSO to the preconditioned response variable is consistent as the number of predictors and observations increases. Moreover, when the observational noise is rather large, the suggested procedure can give a more accurate estimate than LASSO. We illustrate our method on some real problems, including survival analysis with microarray data.

1. Introduction. In this paper, we consider the problem of fitting linear (and other related) models to data for which the number of features p greatly exceeds the number of samples n. This problem occurs frequently in genomics, for example, in microarray studies in which p genes are measured on n biological samples.

The problem of model selection for data where number of variables is typically comparable or much larger than the sample size has received a lot of attention recently. In particular, various penalized regression methods are being widely used as means of selecting the variables having nonzero contribution in a regression model. Among these tools the L_1 penalized regression or LASSO (Tibshirani [16]) is one of the most popular techniques. The Least Angle Regression (LAR) procedure Efron et al. [5] provides a method for fast computation of LASSO solution in regression problems. Osborne, Presnell and Turlach [12] derived the optimality conditions associated with the LASSO solution. Donoho and Elad [4] and Donoho [3] proved some analytical properties of the L_1 penalization approach for deter-

Received December 2006; revised March 2007.
[1]Supported in part by NSF Grant DMS-05-05676 and NIH Grant 2R01 CA 72028-07.
[2]Supported in part by NSF Grant DMS-99-71405 and NIH Contract N01-HV-28183.
AMS 2000 subject classification. 62J07.
Key words and phrases. Model selection, prediction error, lasso.

mining the sparsest solution for an underdetermined linear system. Some statistical properties of the LASSO-based estimator of the regression parameter have been derived by Knight and Fu [9]. In the context of high-dimensional graphs, Meinshausen and Bühlmann [11] showed that the variable selection method based on LASSO can be consistent if the underlying model satisfies some conditions. Various other model selection criteria have been proposed in high-dimensional regression problems. Fan and Li [6] and Shen and Ye [15] gave surveys of some of these methods.

However, when the number of variables (p) is much larger than the number of observations [precisely $p_n \sim cn^{\xi}$ for some $\xi \in (0, 1)$] [10] showed that the convergence rate of risk of the LASSO estimator can be quite slow. For finite-dimensional problems, Zou [20] found a necessary condition for the covariance matrix of the observations, without which the LASSO variable selection approach is inconsistent. Zhao and Yu [19] derived a related result for their $p > N$ case.

Various modifications to LASSO have been proposed to ensure that on one hand, the variable selection process is consistent and on the other, the estimated regression parameter has a fast rate of convergence. Fan and Li [6] proposed the Smoothly Clipped Absolute Deviation (SCAD) penalty for variable selection. Fan and Peng [7] discussed the asymptotic behavior of this and other related penalized likelihood procedures when the dimensionality of the parameter is growing. [20] proposed a nonnegative Garrote-type penalty (that is reweighted by the least squares estimate of the regression parameter) and showed that this estimator has adaptivity properties when p is fixed. Meinshausen [10] proposed a relaxation to the LASSO penalty after initial model selection to address the problem of high bias of LASSO estimate when p is very large.

All of these methods try to solve two problems at once: (1) find a good predictor $\hat{y}$ and, (2) find a (hopefully small) subset of variables to form the basis for this prediction. When $p \gg n$, these problems are especially difficult. In this paper, we suggest that they should be solved separately, rather than both at once. Moreover, the method we propose utilizes the correlation structure of the predictors, unlike most of the methods cited. We propose a two-stage approach:

(a) find a consistent predictor $\hat{y}$ of the true response,

(b) using the *preconditioned* outcome $\hat{y}$, apply a model fitting procedure (such as forward stagewise selection or the LASSO) to the data $(\mathbf{x}, \hat{y})$.

In this paper, we show that the use of $\hat{y}$ in place of y in the model selection step (b) can mitigate the effects of noisy features on the selection process under the setting of a *latent variable model* for the response, when the number of predictor variables that are associated with the response grows at a slower rate than the number of observations, even though the nominal dimension of the predictors can grow at a much faster rate.

This paper is organized as follows. In Section 2, we define the preconditioning method and give an example from a latent variable model. Section 3 discusses a real example from a kidney cancer microarray study, and application of the idea to

other settings such as survival analysis. In Section 4, we give details of the latent variable model, and show that the LASSO applied to the preconditioned response yields a consistent set of predictors, as the number of features and samples goes to infinity. Finally, in Section 5, we discuss and illustrate the preconditioning idea for classification problems.

2. Preconditioning. Suppose that the feature measurements are $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and outcome values y_i, for $i = 1, 2, \ldots, n$. Our basic model has the form

$$\mathbb{E}(y_i|x_i) = \theta_0 + \sum_{j=1}^{p} x_{ij}\theta_j, \qquad i = 1, 2, \ldots, n. \tag{1}$$

Two popular methods for fitting this model are forward stepwise selection (FS) and the LASSO (Tibshirani [16]). The first method successively enters the variable that most reduces the residual sum of squares, while the second minimizes the penalized criterion

$$J(\theta, \mu) = \sum_i \left(y_i - \theta_0 + \sum_{j=1}^{p} \theta_j x_{ij} \right)^2 + \mu \sum_{j=1}^{p} |\theta_j|. \tag{2}$$

Efron et al. [5] develop the least angle regression (LAR) algorithm, for fast computation of the LASSO for all values of the tuning parameter $\mu \geq 0$.

Usually model selection in the general model (1) is quite difficult when $p \gg n$, and our simulations confirm this. To get better results, we may need further assumptions about the underlying model relating y_i to x_i. In this paper, we assume that y_i and x_i are connected via a low-dimensional *latent variable model*, and use a method that we shall refer to as *preconditioning* to carry out model selection. In this approach, we first find a consistent estimate $\hat{y}_i$ by utilizing the latent variable structure, and then apply a fitting procedure such as forward stepwise regression or the LASSO to the data $(x_i, \hat{y}_i)$, $i = 1, 2, \ldots, n$. The main technique that we consider for the initial preconditioning step is supervised principal components (SPC) [1, 2]. This method works as follows:

(a) we select the features whose individual correlation with the outcome is large, and

(b) using just these features, we compute the principal components of the matrix of features, giving $\hat{V}_1, \hat{V}_2, \ldots, \hat{V}_{\min\{N,p\}}$. The prediction $\hat{y}_i$ is the least squares regression of y_i on the first K of these components.

Typically, we use just the first or first few supervised principal components. Bair et al. [1] show that under an assumption about the sparsity of the population principal components, as $p, n \to \infty$, supervised principal components gives consistent estimates for the regression coefficients while the usual principal components regression does not. We give details of this model in Section 4, and provide a simple example next.

2.1. *Example*: *latent variable model*. The following example shows the main idea in this paper. Consider a model of the form:

$$Y = \beta_0 + \beta_1 V + \sigma_1 Z. \tag{3}$$

In addition, we have measurements on a set of features X_j indexed by $j \in \mathcal{A}$, for which

$$X_j = \alpha_{0j} + \alpha_{1j} V + \sigma_0 e_j, \qquad j \in 1, \ldots, p. \tag{4}$$

The quantity V is an unobserved or *latent* variable. The set $\mathcal{A}$ represents the important features (meaning that $\alpha_{1j} \neq 0$, for $j \in \mathcal{A}$) for predicting Y_i. The errors Z_i and e_{ij} are assumed to have mean zero and are independent of all other random variables in their respective models. All random variables (V, Z, e_j) have a standard Gaussian distribution.

2.2. *Example* 1. For illustration, we generated data on $p = 500$ features and $n = 20$ samples, according to this model, with $\beta_1 = 2$, $\beta_0 = 0$, $\alpha_{0j} = 0$, $\alpha_{1j} = 1$, $\sigma_1 = 2.5$, $\mathcal{A} = \{1, 2, \ldots, 20\}$. Our goal is to predict Y from $X_1, X_2, \ldots, X_p$, and in the process, discover the fact that only the first 20 features are relevant. This is a difficult problem. However if we guess (correctly) that the data were generated from model (4), our task is made easier. The left panel of Figure 1 shows the correlations $\mathrm{Corr}(V, X_j)$ plotted versus $\mathrm{Corr}(Y, X_j)$ for each feature j. The first

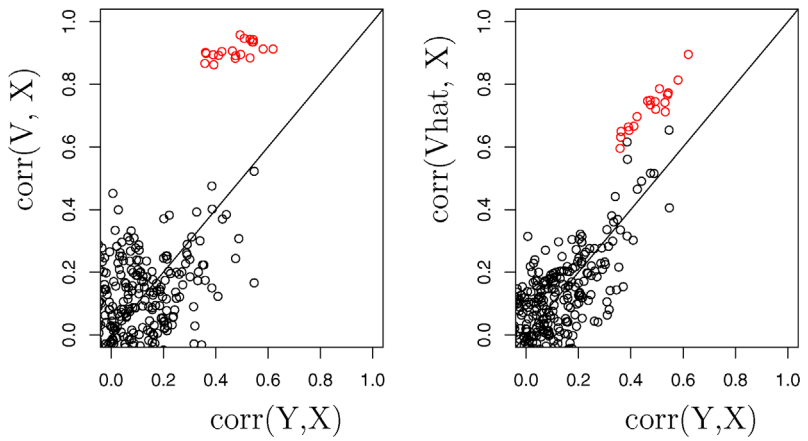

FIG. 1. *Results for simulated data. Left panel shows the correlation between the true latent variable V and gene expression X for each of the genes plotted against the correlation between Y and gene expression. The truly nonnull genes are shown in red. The right panel is the same, except that the estimated latent variable* $\hat{V}$ *(from supervised principal components) replaces V. We see that correlation with either the true or estimated latent factor does a better job at isolating the truly nonnull genes.*

20 features are plotted in red, and can be distinguished much more easily on the basis of $\text{Corr}(V, X_j)$ than $\text{Corr}(Y, X_j)$. However this requires knowledge of the underlying latent factor V, which is not observed.

The right panel shows the result when we instead estimate V_i from the data, using the first supervised principal component. We see that the correlations of each feature with the estimated latent factor also distinguishes the relevant from the irrelevant features.

Not surprisingly, this increased correlation leads to improvements in the performance of selection methods, as shown in Table 1. We applied four selection methods to the 20 simulated data sets from this model: FS: simple forward stepwise regression; SPC/FS: forward stepwise regression applied to the preconditioned outcome from supervised principal components; LASSO and SPC/LASSO: LASSO applied to preconditioned outcome from supervised principal components. The table shows the average number of good variables selected among the first 1, 2, 5, 10 and 20 variables selected and the corresponding test errors. Preconditioning clearly helps both forward selection and the LASSO.

2.3. *Example* 2. The second example was suggested by a referee. It is somewhat artificial but exposes an important assumption that is made by our procedure. We define random variables (Y, X_1, X_2, X_3) having a Gaussian distribution with mean zero and inverse covariance matrix

$$\Sigma^{-1} = \begin{pmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 0 & 1 \\ 1 & 0 & 2 & 1 \\ 1 & 1 & 1 & 2 \end{pmatrix}.$$

We define 297 additional predictors that are $N(0, 1)$. The population regression coefficient is $\beta = (-1, -1, -1, 0, 0, \ldots)$ while the (marginal) correlation of each

TABLE 1

Four selection methods to the 20 *simulated data sets from the model of Example* 1. *Shown are the number of good variables selected among the first* 1, 2, 10 *and* 20 *variables selected and the corresponding test errors. Preconditioning clearly helps in both cases, and the LASSO outperforms forward selection*

	Mean # of good variables when selecting first				**Test error when selecting first**			
Method	**1**	**5**	**10**	**20**	**1**	**5**	**10**	**20**
FS	0.82	0.98	1.12	1.58	267.36	335.4	353.52	357.07
SPC/FS	0.94	2.66	2.86	3.12	241.88	229.47	231.52	232.28
LASSO	0.88	2.05	3.17	3.29	206.54	184.56	186.71	205.85
SPC/LASSO	0.92	4.21	7.75	9.71	212.23	197.07	183.04	178.19

TABLE 2
Performance of LASSO and preconditioned LASSO in the second simulation example

	Mean # of good variables when selecting first			
Method	**1**	**2**	**3**	**4**
LASSO	1.0	2.0	3.0	3.0
SPC/LASSO	1.0	2.0	2.0	2.0

predictor with Y is $\rho = (-0.5, -0.5, 0, 0, 0, \ldots)$. Hence X_3 has zero marginal correlation with Y but has a nonzero partial correlation with Y [since $(\Sigma^{-1})_{14} = 1$]. The number of good variables when selecting the first 1, 2, 3 or 4 predictors is shown in Table 2.

We see that the LASSO enters the 3 good predictors first in every simulation, while the preconditioned version ignores the 3rd predictor. Supervised principal components screens out this predictor, because it is marginally independent of Y.

Preconditioning with supervised principal components assumes that any important predictor (in the sense of having significantly large nonzero regression coefficient) will also have a substantial marginal correlation with the outcome. This need not be true in practice, but we believe it will often be a good working hypothesis in many practical problems.

2.4. *Example* 3. Our third simulation study compares the LASSO to the preconditioned LASSO, in a more neutral setting. We generated 1000 predictors, each having a $N(0, 1)$ distribution marginally. The first 40 predictors had a pairwise correlation of 0.5, while the remainder were uncorrelated.

The outcome was generated as

$$Y = \sum_{j=1}^{40} \beta_j X_j + \sigma Z \tag{5}$$

with $Z, \beta_j \sim N(0, 1)$ and $\sigma = 5$. Hence the outcome is only a function of the first 40 ("good") predictors.

We generated 100 datasets from this model: the average number of good variables selected by the LASSO and preconditioned LASSO is shown in Table 3. Note that with just $n = 50$ samples, the maximum number of predictors in the model is also 50. While neither method is successful at isolating the bulk of the 40 good predictors, the preconditioned LASSO finds twice as many good predictors as the LASSO in the full model.

TABLE 3
Performance of LASSO and preconditioned LASSO in the third simulation example

	Mean # of good variables when selecting first			
Method	**5**	**10**	**20**	**50**
LASSO	2.92	5.88	9.04	9.16
SPC/LASSO	2.49	5.13	10.32	19.73

3. Examples.

3.1. *Kidney cancer data*. Zhao, Tibshirani and Brooks [18] collected gene expression data on 14,814 genes from 177 kidney patients. Survival times (possibly censored) were also measured for each patient, as well as a number of clinical predictors including the grade of the tumor: 1 (good) to 4 (poor).

The data were split into 88 samples to form the training set and the remaining 89 formed the test set. For illustration, in this section, we try to predict grade from gene expression. In the next section we predict survival time (the primary outcome of interest) from gene expression. Figure 2 shows the training and test set correlations between grade and its prediction from different methods. We see that for both forward selection and the LASSO, use of the supervised principal component prediction $\hat{y}$ as the outcome variable (instead of y itself) makes the procedure less greedy in the training set and yields higher correlations in the test set. While the correlations in the test set are not spectacularly high, for SPC/FS and SPC/LASSO they do result in a better prediction in the test set.

3.2. *Application to other regression settings*. Extension of our proposal to other kinds of regression outcomes is very simple. The only change is in step (a) of supervised principal components algorithm, where we replace the correlation by an appropriate measure of association. In particular, the likelihood score statistic is an attractive choice.

3.3. *Survival analysis*. Perhaps the most common version of the $p > n$ regression problem in genomic studies is survival analysis, where the outcome is patient survival (possibly censored). Then we use the partial likelihood score statistic from Cox's proportional hazards score statistic (see Chapter 4 of Kalbfleisch and Prentice [8]), in step (a) of supervised principal components. After that, we can (conveniently) use the usual least squares version of FS or LASSO in step (2) of the modeling process. Hence, the computational advantages of the least angle regression algorithm can be exploited.

Figure 3 shows the result of applying forward stepwise Cox regression (top left panel), forward stepwise selection applied to the SPC predictor (top right panel),

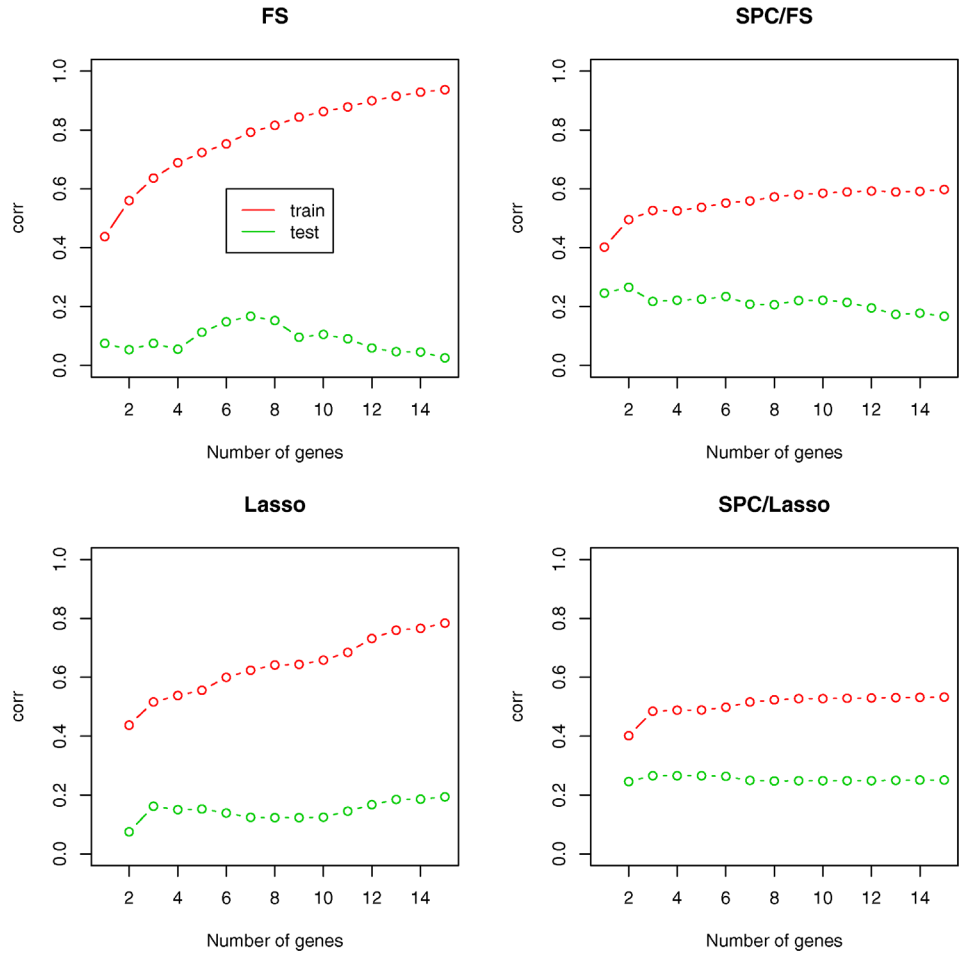

FIG. 2. *Kidney cancer data*: *predicting tumor grade*. *Correlation of different predictors with the true outcome*, *in the training and test sets*, *as more and more genes are entered*.

LASSO for the Cox model (bottom left panel) and LASSO applied to the SPC predictor (bottom right panel). The bottom left panel was computed using the `glmpath` R package of Park and Hastie [13], available in the CRAN collection. In each case, we obtain a predictor $\hat{y}$, and then use $\hat{y}$ as a covariate in a Cox model, in either the training or test set. The resulting p-values from these Cox models are shown in the figure. We see that forward stepwise Cox regression tends to overfit in the training set, and hence, the resulting test-set p-values are not significant. The two stage SPC/FS procedure fits more slowly in the training set, and hence achieves smaller p-values in the test set. "SPC/LASSO," the LASSO applied to the preconditioned response from supervised principal components, performs best and

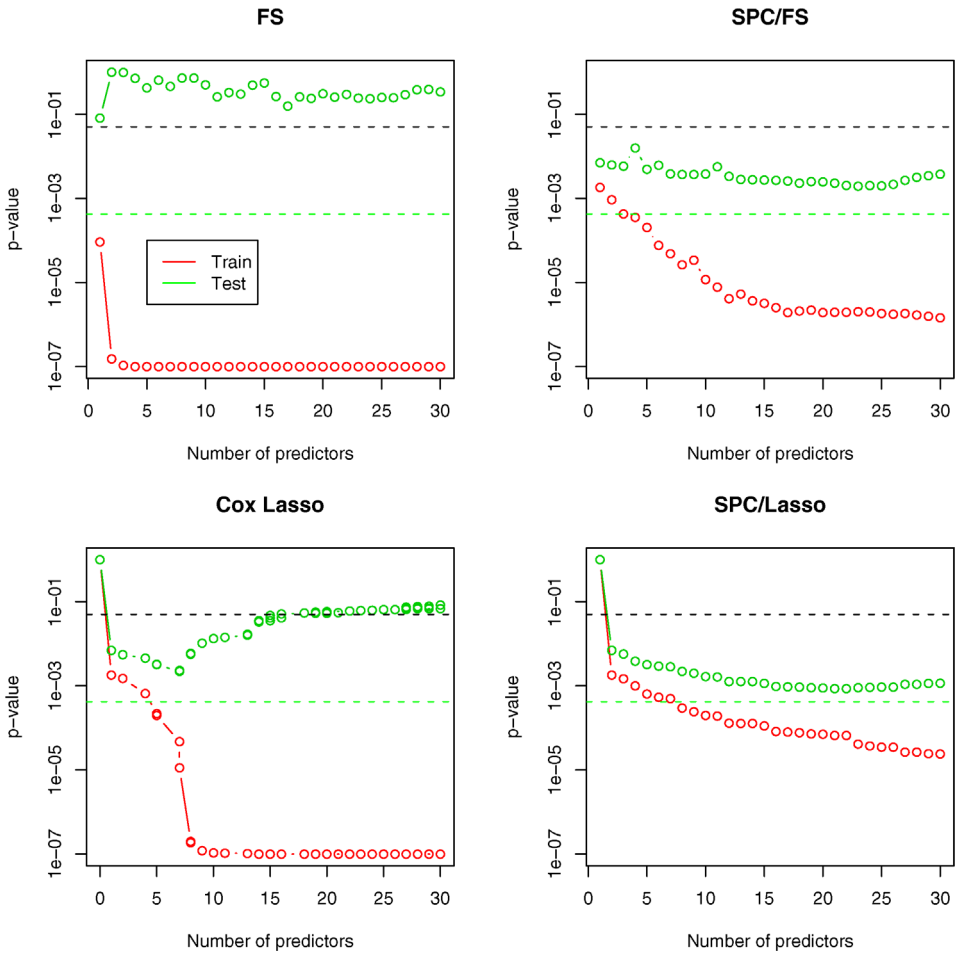

FIG. 3. *Kidney cancer data*: *predicting survival time*. *Training set p-values* (*red*) *and test set p-values* (*green*) *for four different selection methods as more and more genes are entered*. *Horizontal broken lines are drawn at* 0.05 (*black*) *and the test set p-value for the supervised principal component predictor* 0.00042 (*green*).

is also computationally convenient: it uses the fast LAR algorithm for the LASSO, applied to the preconditioned response variable.

The horizontal green line shows the test set p-value of the supervised principal component predictor. We see that the first 10 or 15 genes chosen by the LASSO have captured the signal in this predictor.

We have used the preconditioning procedure in real microarray studies. We have found that it is useful to report to investigators not just the best 10 or 15 gene model, but also any genes that have high correlation with this set. The enlarged set can be useful in understanding the underlying biology in experiment, and also for building assays for future clinical use. A given gene might not be well measured

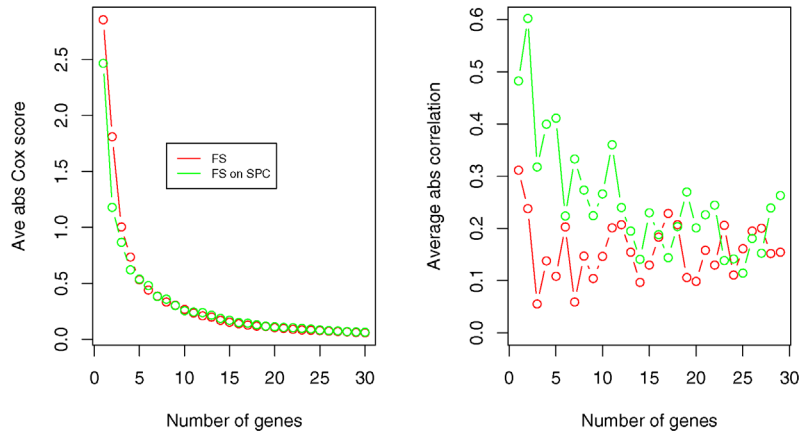

FIG. 4. *Kidney cancer data*: *predicting survival time*. *Left panel shows the average absolute Cox score of the first k genes entered by forward stepwise selection* (*red*) *and the preconditioned version* (*green*), *as k runs from* 1 *to* 30. *The right panel shows the average absolute pairwise correlation of the genes for both methods.*

on a microarray for a variety of reasons, and hence it is useful to identify surrogate genes that may be used in its place.

Figure 4 shows the average absolute Cox score of the first k features entered by forward stepwise selection (red) and the preconditioned version (green), as k runs from 1 to 30. The right panel shows the average absolute pairwise correlation of the genes for both methods. We see that the methods enter features of about the same strength, but preconditioning enters genes that are more highly correlated with one another.

4. Asymptotic analysis. In this section we lay down a mathematical formulation of the problem and preconditioning procedure in the context of a latent factor model for the response. We show that the procedure combining SPC with LASSO, under some assumptions about the correlation structure among the variables, leads to asymptotically consistent variable selection in the Gaussian linear model setting. We consider the class of problems where one observes n independent samples $(y_i, \mathbf{x}_i)$ where y_i is a one-dimensional response and $\mathbf{x}_i$ is a p-dimensional predictor. Individual coordinates of the vector $\mathbf{x}_i$ are denoted by x_{ij} where the index $j \in \{1, \ldots, p\}$ correspond to the jth predictor. We denote the $n \times p$ matrix $((x_{ij}))_{1 \le i \le n, 1 \le j \le p}$ by $\mathbf{X}$ and the vector $(y_i)_{i=1}^n$ by Y. Henceforth, unless otherwise stated, we do not make a distinction between the realized value $(Y, \mathbf{X})$ and the random elements (viz., the response and the p predictors) that they represent.

The interest is in identifying the set of predictors X_j which are (linearly) related to Y. A regression model will be of the form $\mathbb{E}(Y|\mathbf{x}) = \theta^T \mathbf{x}$ for some $\theta \in \mathbb{R}^p$. Here we assume that the joint distribution of $\mathbf{X}$ is Gaussian with zero mean and

covariance matrix $\Sigma \equiv \Sigma_p$. The relationship between Y and $\mathbf{X}$ is assumed to be specified by a latent component model to be described below.

4.1. *Model for* $\mathbf{X}$. Suppose that the spectral decomposition of Σ is given by $\Sigma = \sum_{k=1}^{p} \ell_k \mathbf{u}_k \mathbf{u}_k^T$, where $\ell_1 \geq \cdots \geq \ell_p \geq 0$ and $\mathbf{u}_1, \ldots, \mathbf{u}_p$ form an orthonormal basis of $\mathbb{R}^p$. We consider the following model for Σ.

Assume that there exists an $M \geq 1$ such that

$$\begin{aligned} \ell_k &= \lambda_k + \sigma_0^2, \qquad k = 1, \ldots, M, \quad \text{and} \\ \ell_k &= \sigma_0^2, \qquad\qquad k = M+1, \ldots, p, \end{aligned} \tag{6}$$

where $\lambda_1 \geq \cdots \geq \lambda_M > 0$ and $\sigma_0 > 0$. This model will be referred to as the "noisy factor model." To see this, notice that under the Gaussian assumption the matrix $\mathbf{X}$ can be expressed as

$$\mathbf{X} = \sum_{k=1}^{M} \sqrt{\lambda_k} \mathbf{v}_k \mathbf{u}_k^T + \sigma_0 \mathbf{E}, \tag{7}$$

where $\mathbf{v}_1, \ldots, \mathbf{v}_M$ are i.i.d. $N_n(0, I)$ vectors (the factors), and $\mathbf{E}$ is an $n \times p$ matrix with i.i.d. $N(0, 1)$ entries, and is independent of $\mathbf{v}_1, \ldots, \mathbf{v}_M$. This matrix is viewed as a noise matrix.

In the analysis presented in this paper throughout, we use (7) as the model for $\mathbf{X}$, even though it can be shown that the analysis applies even in the case where $\ell_{K+1}, \ldots, \ell_p$ are decreasing and sufficiently well separated from $\ell_1, \ldots, \ell_K$.

4.2. *Model for* Y. Assume the following regression model for Y. Note that this is a more general version of (3), even though we assume that Y has (unconditional) mean 0.

$$Y = \sum_{k=1}^{K} \beta_k \mathbf{v}_k + \sigma_1 Z, \tag{8}$$

where $\sigma_1 > 0$, $1 \leq K \leq M$, and Z has $N_n(0, I)$ distribution and is independent of $\mathbf{X}$.

4.3. *Least squares and feature selection*. We derive expressions for the marginal correlations between Y and X_j, for $j = 1, \ldots, p$ and the (population) least squares solution, namely $\theta := \arg\min_\zeta \mathbb{E}\|Y - \mathbf{X}\zeta\|_2^2$, in terms of the model parameters. Let $\mathcal{P} := \{1, \ldots, p\}$. The marginal correlation between $\mathbf{X} = (X_j)_{j=1}^p$ and Y is given by

$$\Sigma_{\mathcal{P}y} := (\mathbb{E}(X_j Y))_{j=1}^p = \sum_{k=1}^{K} \beta_k \sqrt{\lambda_k} \mathbf{u}_k. \tag{9}$$

The population regression coefficient of Y on $\mathbf{X}$, is given by

$$\begin{aligned}\theta &= \Sigma^{-1}\Sigma_{\mathcal{P}y}\\ &= \left[\sum_{k=1}^{M}\lambda_k\mathbf{u}_k\mathbf{u}_k^T + \sigma_0^2 I\right]^{-1}\left[\sum_{k=1}^{K}\beta_k\sqrt{\lambda_k}\mathbf{u}_k\right]\\ &= \left[\sum_{k=1}^{M}\frac{1}{\lambda_k+\sigma_0^2}\mathbf{u}_k\mathbf{u}_k^T + \frac{1}{\sigma_0^2}\left(I - \sum_{k=1}^{M}\mathbf{u}_k\mathbf{u}_k^T\right)\right]\left[\sum_{k=1}^{K}\beta_k\sqrt{\lambda_k}\mathbf{u}_k\right] \qquad (10)\\ &= \sum_{k=1}^{K}\beta_k\frac{\sqrt{\lambda_k}}{\lambda_k+\sigma_0^2}\mathbf{u}_k\\ &= \sum_{k=1}^{K}\beta_k\ell_k^{-1}\sqrt{\lambda_k}\mathbf{u}_k.\end{aligned}$$

Now, define $\mathbf{w}_j = (\sqrt{\lambda_1}u_{j1},\ldots,\sqrt{\lambda_K}u_{jK})^T$. Let $\mathcal{D} = \{j : \|\mathbf{w}_j\|_2 \neq 0\}$. Observe that $\Sigma_{jy} = \beta^T\mathbf{w}_j$ and $\theta_j = \beta^T D_K^{-1}\mathbf{w}_j$, where $D_K = \operatorname{diag}(\ell_1,\ldots,\ell_K)$. So if we define $\mathcal{B} := \{j : \Sigma_{jy} \neq 0\}$ and $\mathcal{A} = \{j : \theta_j \neq 0\}$, then $\mathcal{B} \subset \mathcal{D}$ and $\mathcal{A} \subset \mathcal{D}$.

This gives rise to the regression model:

$$Y = \mathbf{X}\theta + \sigma_\varepsilon\varepsilon, \qquad (11)$$

where

$$\begin{aligned}\sigma_\varepsilon^2 &= \sigma_{yy} - \Sigma_{y\mathcal{P}}\Sigma^{-1}\Sigma_{\mathcal{P}y}\\ &= \sigma_1^2 + \sum_{k=1}^{K}\beta_k^2 - \sum_{k=1}^{K}\beta_k^2\frac{\lambda_k}{\lambda_k+\sigma_0^2} \qquad (12)\\ &= \sigma_1^2 + \sigma_0^2\beta^T D_K^{-1}\beta,\end{aligned}$$

and ε has i.i.d. $N(0,1)$ entries and is independent of $\mathbf{X}$.

Note also that, the population partial covariance between Y and $\mathbf{X}_C$ given $\mathbf{X}_\mathcal{D}$ (given by $\Sigma_{yC|\mathcal{D}} := \Sigma_{yC} - \Sigma_{y\mathcal{D}}\Sigma_{\mathcal{D}\mathcal{D}}^{-1}\Sigma_{\mathcal{D}C}$), for any subset $C \subset \mathcal{D}^c$, where $\mathcal{D}^c := \mathcal{P} \setminus \mathcal{D}$, is 0. However, the corresponding statement is not true in general if one replaces $\mathcal{D}$ by either $\mathcal{A}$ or $\mathcal{B}$. Therefore, ideally, one would like to identify $\mathcal{D}$. However, it may not be possible to accomplish this in general when the dimension p grows with the sample size n. Rather, we define the feature selection problem as the problem of identifying $\mathcal{A}$, while the estimation problem is to obtain an estimate of θ from model (11).

Observe that, if either $K = 1$ or $\lambda_1 = \cdots = \lambda_K$, then $\mathcal{A} = \mathcal{B}$. In the former case, we actually have $\mathcal{A} = \mathcal{B} = \mathcal{D}$. In these special cases, the feature selection problem reduces to finding the set $\mathcal{B}$, which may be done (under suitable identifiability

conditions) just by computing the sample marginal correlations between the response and the predictors and selecting those variables (coordinates) for which the marginal correlation exceeds an appropriate threshold. The major assumptions that we shall make here for solving the problem are that (i) $\mathcal{A} \subset \mathcal{B}$, (ii) $\mathcal{B}$ can be identified from the data (at least asymptotically), (iii) cardinality of $\mathcal{B}$ (and hence that of $\mathcal{A}$) is small compared to n, and (iv) the contribution of the coordinates $\mathcal{B}^c$ in the vectors $\mathbf{u}_1, \ldots, \mathbf{u}_K$ is asymptotically negligible in an L^2 sense. If these conditions are satisfied, then it will allow for the identification of $\mathcal{A}$, even as dimension increases with the sample size. We make these (and other) conditions more precise in Section 4.7.

4.4. *SPC as a preconditioner.* The formulation in the previous section indicates that one may use some penalized regression methods to estimate the regression parameter θ from the model (11). However, standard methods like LASSO do not use the covariance structure of the data. Therefore, if one uses the underlying structure for Σ, and has good estimates of the parameters $(\mathbf{u}_k, \ell_k)$, then one can hope to be able to obtain a better estimate θ, as well as identify $\mathcal{A}$ as $n \to \infty$.

We focus on (7) and (8). In general, it is not possible to eliminate the contribution of $\mathbf{E}$ entirely from an estimate of $\mathbf{v}_k$, even if we had perfect knowledge of $(\mathbf{u}_k, \ell_k)$. To understand this, note that, the conditional distribution of $\mathbf{v}_k$ given $\mathbf{X}$ is the same as the conditional distribution of $\mathbf{v}_k$ given $\mathbf{X}\mathbf{u}_k$. The latter distribution is normal with mean $\frac{\sqrt{\lambda_k}}{\ell_k}\mathbf{X}\mathbf{u}_k$ and covariance matrix $\frac{\sigma_0^2}{\ell_k} I_n$. This means that any reasonable procedure that estimates the parameters $(\mathbf{u}_k, \ell_k)$ can only hope to reduce the effect of the measurement noise in Y, namely $\sigma_1 Z$.

Keeping these considerations in mind, we employ a two stage procedure described in the following section for estimating θ. In order to fit the model (11) using SPC procedure, it is necessary to estimate the eigenvectors $\mathbf{u}_k$, $k = 1, \ldots, M$. When $\frac{p}{n}$ is large (in the sense that the fraction does not converge to 0 as $n \to \infty$), in general it is not possible to estimate $\mathbf{u}_k$ consistently. However, if $\mathbf{u}_k$ are sparse, in the sense of having say q nonzero components, where $\frac{q}{n} \to 0$, then [1] showed that under suitable identifiability conditions, it is possible to get asymptotically consistent estimators of $\mathbf{u}_1, \ldots, \mathbf{u}_K$, where the consistency is measured in terms of convergence of the L^2 distance between the parameter and its estimator.

4.5. *Algorithm.* In this section, we present the algorithm in detail.

Step 1 Estimate $(\mathbf{u}_1, \ell_1), \ldots, (\mathbf{u}_K, \ell_K)$ by SPC procedure in which only those predictors X_j whose empirical correlation with response Y is above a threshold τ_n are used in the eigen-analysis. Call these estimates $\{\widetilde{\mathbf{u}}_k, \widetilde{\ell}_k\}_{k=1}^K$.

Step 2 Let $\widetilde{P}_K := Proj(\widehat{V}_1, \ldots, \widehat{V}_K)$ be the projection onto $\widehat{V}_1, \ldots, \widehat{V}_K$, where $\widehat{V}_k := \frac{1}{\sqrt{\widetilde{\ell}_k}}\mathbf{X}\widetilde{\mathbf{u}}_k$ is the kth principal component of the predictors (under the SPC procedure). Define $\widetilde{Y} = \widetilde{P}_K Y$.

Step 3 Estimate θ from the linear model $\widetilde{Y} = \mathbf{X}\theta +$ error, using the LASSO approach with penalty $\mu_n > 0$.

Since by definition $\frac{1}{n}\langle \mathbf{X}\widetilde{\mathbf{u}}_k, \mathbf{X}\widetilde{\mathbf{u}}_{k'}\rangle = \widetilde{\ell}_k \delta_{kk'}$, it follows that

$$\begin{aligned}\widetilde{P}_K &= Proj(\mathbf{X}\widetilde{\mathbf{u}}_1, \ldots, \mathbf{X}\widetilde{\mathbf{u}}_K)\\ &= \sum_{k=1}^{K} \frac{1}{\|\mathbf{X}\widetilde{\mathbf{u}}_k\|^2}(\mathbf{X}\widetilde{\mathbf{u}}_k)(\mathbf{X}\widetilde{\mathbf{u}}_k)^T \qquad (13)\\ &= \sum_{k=1}^{K} \frac{1}{\widetilde{\ell}_k}\frac{1}{n}(\mathbf{X}\widetilde{\mathbf{u}}_k)(\mathbf{X}\widetilde{\mathbf{u}}_k)^T.\end{aligned}$$

4.6. *Analysis of the projection.* We present an expansion of the projected response $\widetilde{Y} := \widetilde{P}_K Y$ that will be useful for all the asymptotic analyses that follow. Using the representation of $\widetilde{P}_K$ in (13) and invoking (7) and (8), we get

$$\begin{aligned}\widetilde{Y} &= \sum_{k=1}^{K} \frac{\beta_k}{\widetilde{\ell}_k}\frac{1}{n}\langle \mathbf{X}\widetilde{\mathbf{u}}_k, \mathbf{v}_k\rangle \mathbf{X}\widetilde{\mathbf{u}}_k\\ &\quad + \sum_{k=1}^{K}\sum_{k'\neq k}^{K} \frac{\beta_{k'}}{\widetilde{\ell}_k}\frac{1}{n}\langle \mathbf{X}\widetilde{\mathbf{u}}_k, \mathbf{v}_{k'}\rangle \mathbf{X}\widetilde{\mathbf{u}}_k\\ &\quad + \sigma_1 \sum_{k=1}^{K} \frac{1}{\widetilde{\ell}_k}\frac{1}{n}\langle \mathbf{X}\widetilde{\mathbf{u}}_k, Z\rangle \mathbf{X}\widetilde{\mathbf{u}}_k\\ &= \sum_{k=1}^{K} \frac{\beta_k\sqrt{\lambda_k}}{\widetilde{\ell}_k}\frac{1}{n}\|\mathbf{v}_k\|^2\langle \mathbf{u}_k, \widetilde{\mathbf{u}}_k\rangle \mathbf{X}\widetilde{\mathbf{u}}_k\\ &\quad + \sum_{k=1}^{K}\sum_{l\neq k}^{M} \frac{\beta_k\sqrt{\lambda_l}}{\widetilde{\ell}_k}\frac{1}{n}\langle \mathbf{v}_l, \mathbf{v}_k\rangle\langle \mathbf{u}_l, \widetilde{\mathbf{u}}_k\rangle \mathbf{X}\widetilde{\mathbf{u}}_k\\ &\quad + \sum_{k=1}^{K}\sum_{k'\neq k}^{K}\sum_{l=1}^{M} \frac{\beta_{k'}\sqrt{\lambda_l}}{\widetilde{\ell}_k}\frac{1}{n}\langle \mathbf{v}_l, \mathbf{v}_{k'}\rangle\langle \mathbf{u}_l, \widetilde{\mathbf{u}}_k\rangle \mathbf{X}\widetilde{\mathbf{u}}_k\\ &\quad + \sigma_0 \sum_{k=1}^{K}\sum_{k'=1}^{K} \frac{\beta_{k'}}{\widetilde{\ell}_k}\frac{1}{n}\langle \mathbf{E}\widetilde{\mathbf{u}}_k, \mathbf{v}_{k'}\rangle \mathbf{X}\widetilde{\mathbf{u}}_k\\ &\quad + \sigma_1 \sum_{k=1}^{K} \frac{1}{\widetilde{\ell}_k}\frac{1}{n}\langle \mathbf{X}\widetilde{\mathbf{u}}_k, Z\rangle \mathbf{X}\widetilde{\mathbf{u}}_k \qquad (14)\end{aligned}$$

$$
\begin{aligned}
&= \mathbf{X}\theta + \mathbf{X}\sum_{k=1}^{K} \beta_k \sqrt{\lambda_k}\left(\frac{1}{\widetilde{\ell}_k}\frac{\|\mathbf{v}_k\|^2}{n}\langle \mathbf{u}_k, \widetilde{\mathbf{u}}_k\rangle \widetilde{\mathbf{u}}_k - \frac{1}{\ell_k}\mathbf{u}_k\right) \\
&\quad + \sum_{k=1}^{K}\sum_{k'\neq k}^{K} \frac{\beta_{k'}\sqrt{\lambda_{k'}}}{\widetilde{\ell}_k}\frac{\|\mathbf{v}_{k'}\|^2}{n}\langle \mathbf{u}_{k'}, \widetilde{\mathbf{u}}_k\rangle \mathbf{X}\widetilde{\mathbf{u}}_k \\
&\quad + \sum_{k=1}^{K}\sum_{k'\neq k}^{K} \frac{\beta_{k'}\sqrt{\lambda_k}}{\widetilde{\ell}_k}\frac{1}{n}\langle \mathbf{v}_k, \mathbf{v}_{k'}\rangle\langle \mathbf{u}_k, \widetilde{\mathbf{u}}_k\rangle \mathbf{X}\widetilde{\mathbf{u}}_k \\
&\quad + \sigma_0 \sum_{k=1}^{K}\sum_{k'=1}^{K} \frac{\beta_{k'}}{\widetilde{\ell}_k}\frac{1}{n}\langle \mathbf{E}\widetilde{\mathbf{u}}_k, \mathbf{v}_{k'}\rangle \mathbf{X}\widetilde{\mathbf{u}}_k \\
&\quad + \sigma_1 \sum_{k=1}^{K} \frac{1}{\widetilde{\ell}_k}\frac{1}{n}\langle \mathbf{X}\widetilde{\mathbf{u}}_k, Z\rangle \mathbf{X}\widetilde{\mathbf{u}}_k + R_n,
\end{aligned}
$$

for some vector $R_n \in \mathbb{R}^n$. This is an asymptotically unbiased regression model for estimating θ provided $(\widetilde{\mathbf{u}}_k, \widetilde{\ell}_k)_{k=1}^K$ is an asymptotically consistent estimator for $(\mathbf{u}_k, \ell_k)_{k=1}^K$.

4.7. *Assumptions*. In this section we give sufficient conditions for the consistency of the variable selection aspect of the SPC preconditioning procedure. The methods of Zou [20] and Knight and Fu [9] are not applicable in our situation since the dimension is growing with the sample size. For most parts, we make assumptions similar to those in Meinshausen and Bühlmann [11] for the relationship among the variables.

A1 The eigenvalues $\lambda_1, \ldots, \lambda_M$ satisfy:

(i) $\lambda_1 > \cdots > \lambda_K > \lambda_{K+1} \geq \cdots \geq \lambda_M \geq 0$.
(ii) $\min_{1\leq k\leq K}(\lambda_k - \lambda_{k+1}) \geq C_0$ for some $C_0 > 0$ (fixed).
(iii) $\lambda_1 \leq \Lambda_{\max}$ for some $\Lambda_{\max}$ fixed. Also, σ_0 is fixed.

A2 $\sigma_1^2 = O(n^{\kappa_0})$ for some $\kappa_0 \in (0, \frac{1}{2})$.

A3 $|\mathcal{A}| = q_n$, $|\mathcal{B}| = \overline{q}_n$ such that $\overline{q}_n = O(n^{\kappa_1})$ for some $\kappa_1 \in (0, \frac{1}{2})$.

A3′ p_n, the number of variables, satisfies the condition that there is an $\alpha > 0$ such that $\log p_n = O(n^{\alpha})$ for some $\alpha \in (0, 1)$.

A4 There exists a ρ_n satisfying $\rho_n n^{1/2}(\log p_n)^{-1/2} \to \infty$ as $n \to \infty$ such that

$$
\min_{j\in\mathcal{B}} \left|\frac{\Sigma_{jy}}{\sqrt{\Sigma_{jj}\sigma_{yy}}}\right| \geq \rho_n. \tag{15}
$$

A5 There exists a $\overline{\delta}_n$ with $\overline{\delta}_n = o(\frac{\overline{q}_n}{n\log n})$ such that $\sum_{j\notin\mathcal{B}} \|\mathbf{w}_j\|_2^2 \leq \overline{\delta}_n$.

A6 There exists an $\eta_n > 0$ satisfying $\eta_n^{-1} = O(n^{\kappa_2})$ for some $\kappa_2 < \frac{1}{2}(1 - \kappa_0 \vee \kappa_1)$, such that

$$\min_{j \in \mathcal{A}} |\theta_j| \geq \eta_n. \tag{16}$$

A7 There exists a $\delta \in (0, 1)$ such that

$$\|\Sigma_{\mathcal{A}^c \mathcal{A}} \Sigma_{\mathcal{A}\mathcal{A}}^{-1} \operatorname{sign}(\theta_{\mathcal{A}})\|_\infty < \delta. \tag{17}$$

A8 There is a $\vartheta < \infty$ such that,

$$\max_{j \in \mathcal{A}} \|\Sigma_{\mathcal{A}_j \mathcal{A}_j}^{-1} \Sigma_{\mathcal{A}_j j}\|_1 < \vartheta \qquad \text{where } \mathcal{A}_j := \mathcal{A} \setminus \{j\}. \tag{18}$$

A few remarks about these conditions are in order. First, condition A1 about the separation of the eigenvalues is not really necessary, but is assumed to avoid the issue of unidentifiability of an eigenvector. However, the scaling of the eigenvalues is important for the analysis. We remark that it is not necessary that the eigenvalues $\lambda_1, \ldots, \lambda_M$ are the M largest eigenvalues of Σ in order for the conclusions to hold. All that is necessary is that these are the leading eigenvalues of the matrix $\Sigma_{\mathcal{D}\mathcal{D}}$, and there is enough separation from the other eigenvalues of Σ. However, this assumption is made to simplify the exposition.

Next, the condition that $\overline{q_n} = o(n)$ (implicit from condition A3) is necessary for the consistency of the estimated eigenvectors $\widetilde{\mathbf{u}}_k$ from supervised PCA. Condition A4 is necessary for the identifiability of the set $\mathcal{B}$. Condition A5 implies that the contribution of the predictors $\{X_j : j \in \mathcal{D} \setminus \mathcal{B}\}$ is negligible in our analysis. Note that $\overline{\delta_n}$ is essentially measuring the "selection bias" for restricting analysis to $\mathcal{B}$ rather than $\mathcal{D}$. Again, the assumption about the rate of decay of $\overline{\delta_n}$ can be relaxed at the cost of more involved analysis and smaller range of values for μ_n (see also the remark following Corollary 1). Too large a value of $\overline{\delta_n}$ may mean that we may not be able to select the variables consistently. Condition A6 is an identifiability condition for set $\mathcal{A}$.

Condition A7 is needed to guarantee consistency of the variable selection by LASSO after projection. This condition was shown to be necessary for variable selection in finite-dimensional LASSO regression by Zou [20] and also, implicitly by Meinshausen and Bühlmann [11]. Zhao and Yu [19] termed this the "irrepresentable condition" and showed that it is nearly necessary and sufficient for consistency of model selection by LASSO when $p, n \to \infty$. A sufficient condition for this to hold is that $\max_{j \in \mathcal{A}^c} \|\Sigma_{\mathcal{A}\mathcal{A}}^{-1} \Sigma_{\mathcal{A}j}\|_1 < \delta$. Observe that $\Sigma_{\mathcal{A}\mathcal{A}}^{-1} \Sigma_{\mathcal{A}j}$ is the population regression coefficient in the regression of X_j on $\{X_l : l \in \mathcal{A}\}$. If we are using the estimate $\widehat{\theta}^{\widehat{\mathcal{B}},\mu}$ then (see proof of Lemma 2) we can replace A7 by the weaker requirement

$$\|\Sigma_{\mathcal{A}^c \cap \mathcal{B}, \mathcal{A}} \Sigma_{\mathcal{A}\mathcal{A}}^{-1} \operatorname{sign}(\theta_{\mathcal{A}})\|_\infty < \delta \qquad \text{for some } \delta \in (0, 1).$$

4.8. *LASSO solution.* We use the symbol μ to denote the penalty parameter in LASSO. The LASSO estimate of θ, after preconditioning, is given by

$$\widehat{\theta}^{\mu} = \arg\min_{\zeta \in \mathbb{R}^p} \frac{1}{n} \|\widetilde{Y} - \mathbf{X}\zeta\|_2^2 + \mu \|\zeta\|_1. \tag{19}$$

We also define the *selected LASSO estimate* of θ by

$$\widehat{\theta}^{\widehat{\mathcal{B}},\mu} = \arg\min_{\zeta \in \mathbb{R}^p, \zeta_{\widehat{\mathcal{B}}^c}=0} \frac{1}{n} \|\widetilde{Y} - \mathbf{X}\zeta\|_2^2 + \mu \|\zeta\|_1. \tag{20}$$

For future use, we define the *restricted LASSO estimate* of θ to be

$$\widehat{\theta}^{\mathcal{A},\mu} = \arg\min_{\zeta \in \mathbb{R}^p, \zeta_{\mathcal{A}^c}=0} \frac{1}{n} \|\widetilde{Y} - \mathbf{X}\zeta\|_2^2 + \mu \|\zeta\|_1. \tag{21}$$

The notation used here follow Meinshausen and Bühlmann [11].

4.9. *Consistency of variable selection.* We shall prove most of our consistency results for the estimate $\widehat{\theta}^{\widehat{\mathcal{B}},\mu}$ and indicate how (and under what conditions) the same may be proved for the unrestricted estimator $\widehat{\theta}^{\mu}$. As we shall see, when the model assumptions hold the former estimator is more reliable under a wider range of possible dimensions. The latter can consistently select the model essentially when $p_n = O(n^{\kappa})$ for some $\kappa < \infty$. In order to prove these results, it will be convenient for us to assume that we have two independent subsamples of size n each, so that the total sample size is $2n$. And we also assume that Step 1 of the variable selection algorithm (estimating $\mathcal{B}$) is performed on the first subsample and the other steps are performed on the second subsample. This extra assumption simplifies our proofs (see the proof of Proposition 4 in the Appendix) somewhat. Further, we shall assume that K, the number of latent components for response Y, is known. The results presented here hold uniformly w.r.t. the parameters satisfying assumptions A1–A8.

Let $\widehat{\mathcal{A}}_{\widehat{\mathcal{B}},\mu}$ (resp. $\widehat{\mathcal{A}}_{\mu}$) denote the set of nonzero coordinates of the vector $\widehat{\theta}^{\widehat{\mathcal{B}},\mu}$ (resp. $\widehat{\theta}^{\mu}$). Whenever the context is clear, we shall drop the subscripts from $\widehat{\mathcal{A}}$. In the following, ζ will be used to denote a generic value of the parameter.

PROPOSITION 1. *Let $\widehat{\mathcal{B}}$ denote the set of coordinates selected by the preliminary thresholding scheme of SPC with threshold τ_n. Given any $c_1 > 1$, and there is a $\tau_n(c_1) := d_1\sqrt{\frac{\log p_n}{n}}$, for some constant $d_1 > 2$, such that, for $n \geq n_{c_1}$,*

$$\mathbb{P}(\widehat{\mathcal{B}} = \mathcal{B}) \geq 1 - n^{-c_1}. \tag{22}$$

Proposition 1 tells us that we can restrict our analysis to the set $\mathcal{B}$ while an-

alyzing the effect of preconditioning and studying the estimator $\widehat{\theta}^{\widehat{\mathcal{B}},\mu}$. Our next result is about the behavior of the estimated eigenvalues and eigenvectors of the matrix $\mathbf{S}_{\widehat{\mathcal{B}}\widehat{\mathcal{B}}} := \frac{1}{n}\mathbf{X}_{\widehat{\mathcal{B}}}^T\mathbf{X}_{\widehat{\mathcal{B}}}$. This result can be proved along the lines of Theorem 3.2 in Paul [14] (see also Bair et al. [1]) and is omitted.

PROPOSITION 2. *Let $(\overline{\mathbf{u}}_{\mathcal{B}k}, \overline{\ell}_k)_{k=1}^K$ denote the first k eigenvector–eigenvalue pairs of $\Sigma_{\mathcal{B}\mathcal{B}}$. Suppose that assumptions* A1–A5 *hold. Then there are functions $\gamma_i = \gamma_i(\lambda_1/\sigma_0, \ldots, \lambda_M/\sigma_0)$, $i = 1, 2$, such that, given $c_2 > 0$ there exist $d_2, d_2' \geq 1$ so that,*

$$\mathbb{P}\Big(\max_{1\leq k\leq K} \|\widetilde{\mathbf{u}}_{\mathcal{B}k} - \overline{\mathbf{u}}_{\mathcal{B}k}\|_2 > d_2\sigma_0\gamma_1\sqrt{\frac{\overline{q}_n \vee \log n}{n}}\Big(1 + \sqrt{\frac{\overline{q}_n \log n}{n}}\Big), \widehat{\mathcal{B}} = \mathcal{B}\Big) = O(n^{-c}),$$

$$\mathbb{P}\Big(\max_{1\leq k\leq K} |\widetilde{\ell}_k - \overline{\ell}_k| > d_2'\sigma_0^2\gamma_2\Big(\sqrt{\frac{\log n}{n}} + \frac{\overline{q}_n \log n}{n}\Big), \widehat{\mathcal{B}} = \mathcal{B}\Big) = O(n^{-c}).$$

THEOREM 1. *Suppose that assumptions* A1–A8 *hold. If $\mu = \mu_n$ satisfies $\mu_n = o(n^{-\kappa_2})$ and $\mu_n n^{(1/2)(1-\kappa_0\vee\kappa_1)} \to \infty$ as $n \to \infty$, then there exists some $c > 1$ such that, for large enough n,*

$$\mathbb{P}(\widehat{\mathcal{A}} \subset \mathcal{A}) \geq 1 - O(n^{-c}), \tag{23}$$

where $\widehat{\mathcal{A}} = \widehat{\mathcal{A}}_{\widehat{\mathcal{B}},\mu_n}$. If moreover, p_n is such that $\frac{q_n \log p_n}{n} = o(1)$ as $n \to \infty$, then (23) *holds with $\widehat{\mathcal{A}} = \widehat{\mathcal{A}}_{\mu_n}$.*

THEOREM 2. *With $\mu = \mu_n$ and $\widehat{\mathcal{A}}$ as in Theorem* 1*, there exists $c > 1$ such that,*

$$\mathbb{P}(\mathcal{A} \subset \widehat{\mathcal{A}}) \geq 1 - O(n^{-c}). \tag{24}$$

Clearly, Theorems 1 and 2 together imply that the SPC/LASSO procedure asymptotically selects the correct set of predictors under the stated assumptions. The proofs of these critically rely on the following three results.

LEMMA 1. *Given $\theta \in \mathbb{R}^p$, let $G(\theta)$ be the vectors whose components are defined by*

$$G_j(\theta) = -\frac{2}{n}\langle \widetilde{Y} - \mathbf{X}\theta, X_j\rangle. \tag{25}$$

A vector $\widehat{\theta}$ *with* $\widehat{\theta}_j = 0$ *for all* $j \in \mathcal{A}^c$ *is a solution of* (21) *if and only if, for all* $j \in \mathcal{A}$,

$$
\begin{aligned}
G_j(\widehat{\theta}) &= -\operatorname{sign}(\widehat{\theta}_j)\mu \qquad && \text{if } \widehat{\theta}_j \neq 0,\\
|G_j(\widehat{\theta})| &\leq \mu && \text{if } \widehat{\theta}_j = 0.
\end{aligned} \tag{26}
$$

Moreover, if the solution is not unique and $|G_j(\widehat{\theta})| < \mu$ *for some solution* $\widehat{\theta}$, *then* $\widehat{\theta}_j = 0$ *for all solutions of* (21).

PROPOSITION 3. *Let* $\widehat{\theta}^{\mathcal{A},\mu}$ *be defined as in* (21). *Then, under the assumptions of Theorem* 1, *for any constant* $c_3 > 1$, *for large enough* n,

$$
\mathbb{P}\big(\operatorname{sign}(\widehat{\theta}_j^{\mathcal{A},\mu_n}) = \operatorname{sign}(\theta_j),\ \text{for all } j \in \mathcal{A}\big) \geq 1 - O(n^{-c_3}). \tag{27}
$$

LEMMA 2. *Define*

$$
\mathcal{E}_{\mathcal{B},\mu} = \Big\{ \max_{j \in \mathcal{A}^c \cap \mathcal{B}} |G_j(\widehat{\theta}^{\mathcal{A},\mu})| < \mu \Big\} \cap \{\widehat{\mathcal{B}} = \mathcal{B}\}. \tag{28}
$$

On $\mathcal{E}_{\mathcal{B},\mu}$, $\widehat{\theta}^{\mathcal{B},\mu}$ *is the unique solution of* (20) *and* $\widehat{\theta}^{\mathcal{A},\mu}$ *is the unique solution of* (21), *and* $\widehat{\theta}^{\widehat{\mathcal{B}},\mu} = \widehat{\theta}^{\mathcal{A},\mu}$. *Also, under the assumptions of Theorem* 1, *there exists a* $c_4 > 1$ *such that, for large enough* n,

$$
\mathbb{P}(\mathcal{E}^c_{\mathcal{B},\mu}) = O(n^{-c_4}). \tag{29}
$$

Further, if we define

$$
\mathcal{E}_{\mu} = \Big\{ \max_{j \in \mathcal{A}^c} |G_j(\widehat{\theta}^{\mathcal{A},\mu})| < \mu \Big\} \cap \{\widehat{\mathcal{B}} = \mathcal{B}\}, \tag{30}
$$

then under the extra assumption that $\frac{q_n \log p_n}{n} = o(1)$, (29) *holds with* $\mathcal{E}_{\mathcal{B},\mu}$ *replaced by* $\mathcal{E}_{\mu}$. *On* $\mathcal{E}_{\mu}$, $\widehat{\theta}^{\mu}$ *is the unique solution of* (19) *and* $\widehat{\theta}^{\mu} = \widehat{\theta}^{\widehat{\mathcal{B}},\mu} = \widehat{\theta}^{\mathcal{A},\mu}$.

4.10. *Effect of projection*. An important consequence of the projection is that the measurement noise Z is projected onto a K dimensional space (that under our assumptions also contains the important components of the predictors of Y). This results in a stable behavior of the residual of the projected response Δ given by

$$
\Delta := \widetilde{Y} - \mathbf{X}\theta = \widetilde{Y} - \mathbf{X}_{\mathcal{A}}\theta_{\mathcal{A}} \tag{31}
$$

even as dimension p_n becomes large. This can be stated formally in the following proposition.

PROPOSITION 4. *Suppose that assumptions* A1–A5 *hold. Then there is a constant* $\gamma_3 := \gamma_3(\sigma_0, \lambda_1, \ldots, \lambda_K + 1)$, *such that for any* $c_6 > 1$ *there exists a constant* $d_6 > 0$ *so that, for large enough* n,

$$
\mathbb{P}\big(\|\Delta\|_2 \leq d_6\big(\gamma_3\sqrt{q_n} \vee \log n + \sigma_1\sqrt{K \log n}\big)\big) \geq 1 - n^{-c_6}. \tag{32}
$$

As a direct corollary to this we have the following result about the risk behavior of the OLS-estimator (under L^2 loss) of the preconditioned data after we have selected the variables by solving the optimization problem (20).

COROLLARY 1. *Suppose that conditions of Theorem 1 hold. Then for any* $c_7 \geq 1$, *there is* $d_7 > 0$ *such that*

$$\mathbb{P}\left(\|\widehat{\theta}^{\widehat{\mathcal{A}}_{\widehat{\mathcal{B}},\mu},\mathrm{OLS}} - \theta\|_2 \leq d_7\sigma_0^{-1}\left(\gamma_3\sqrt{\frac{\overline{q}_n \vee \log n}{n}} + \sigma_1\sqrt{\frac{K\log n}{n}}\right)\right) \geq 1 - n^{-c_7}, \tag{33}$$

where $\widehat{\theta}^{\widehat{\mathcal{A}}_{\widehat{\mathcal{B}},\mu},\mathrm{OLS}} = (\mathbf{X}_{\widehat{\mathcal{A}}}^T\mathbf{X}_{\widehat{\mathcal{A}}})^{-1}\mathbf{X}_{\widehat{\mathcal{A}}}^T\widetilde{Y}$ *and* $\widehat{\mathcal{A}} = \widehat{\mathcal{A}}_{\widehat{\mathcal{B}},\mu_n} = \{j \in \mathcal{P} : \widehat{\theta}^{\widehat{\mathcal{B}},\mu_n} \neq 0\}$.

As a comparison we can think of the situation when $\mathcal{A}$ is actually known, and consider the L^2 risk behavior of the OLS estimator restricted only to the subset of variables $\mathcal{A}$. Then $\widehat{\theta}^{\mathcal{A},\mathrm{OLS}} = (\mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{X}_{\mathcal{A}}^T Y$. Using the fact that conditional on $\mathbf{X}_{\mathcal{A}}$, $\widehat{\theta}_{\mathcal{A}}^{\mathcal{A},\mathrm{OLS}}$ has $N(\theta_{\mathcal{A}}, \sigma_\varepsilon^2(\mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}})^{-1})$ distribution, and the fact that the smallest eigenvalue of $\Sigma_{\mathcal{A},\mathcal{A}}^{-1}$ is at least ℓ_1^{-1}, it follows (using Lemma A.1) that there is a constant $d_7' > 0$ such that

$$\mathbb{P}\left(\|\widehat{\theta}^{\mathcal{A},\mathrm{OLS}} - \theta\|_2 \geq d_7'\ell_1^{-1/2}\sigma_\varepsilon\sqrt{\frac{q_n}{n}}\right) \geq 1 - n^{-c_7}. \tag{34}$$

Comparing (34) with (33), we see that if $q_n \gg \log n$ and $\sigma_1 \gg \sqrt{\overline{q}_n/q_n}$, the estimator $\widehat{\theta}^{\widehat{\mathcal{A}}_{\widehat{\mathcal{B}},\mu},\mathrm{OLS}}$ has better risk performance than $\widehat{\theta}^{\mathcal{A},\mathrm{OLS}}$.

As a remark, we point out that the bound in (33) can be improved under specific circumstances (e.g., when $\overline{\delta}_n$, the "selection bias" term defined in A5, is of a smaller order) by carrying out a second order analysis of the eigenvectors $\{\widetilde{\mathbf{u}}_k\}_{k=1}^K$ (see Appendix of Bair et al. [1]). The same holds for the bounds on the partial correlations $\frac{1}{n}\langle(I - P_{\mathbf{X}_{\mathcal{A}}})X_j, \widetilde{Y}\rangle$, for $j \in \mathcal{A}^c$, given the "signal" variables $\{X_l : l \in \mathcal{A}\}$, that are needed in the proof of Proposition 3 and Lemma 2. However, the result is given here just to emphasize the point that preconditioning stabilizes the fluctuation in $\widetilde{Y} - \mathbf{X}\theta$, and so, partly to keep the exposition brief, we do not present the somewhat tedious and technical work needed to carry out such an analysis.

As a further comparison, we consider the contribution of the measurement noise Z in the maximal empirical partial correlation $\max_{j\in\mathcal{A}^c}|\frac{1}{n}\langle(I - P_{\mathbf{X}_{\mathcal{A}}})X_j, \widetilde{Y}\rangle|$, given $\{X_l : l \in \mathcal{A}\}$. For the preconditioned response, this contribution is [with probability at least $1 - O(n^{-c})$ for some $c > 1$] of the order $O(\frac{\sigma_1\sqrt{\log n}}{\sqrt{n}})$, instead of $O(\frac{\sigma_1\sqrt{\log p_n}}{\sqrt{n}})$ as would be the case if one uses Y instead of $\widetilde{Y}$. So, if $\log p_n \gg \log n$, then the contribution is smaller for the preconditioned response. Formalizing this argument, we derive the following asymptotic result about the model selection property of LASSO estimator that clearly indicates that under latter circumstances

SPC + LASSO procedure can outperform conventional LASSO in terms of variable selection.

PROPOSITION 5. *Suppose that* $\log p_n = cn^\alpha$ *for some* $\alpha \in (0,1)$ *and some* $c > 0$. *Suppose that* $\mathcal{A} = \mathcal{A}_+ \cup \mathcal{A}_-$, *with* $\mathcal{A}_+$ *and* $\mathcal{A}_-$ *disjoint and* $\mathcal{A}_-$ *is nonempty such that* $\|\theta_{\mathcal{A}_-}\|_2 = o(n^{-(1-\alpha)/2})$. *Assume that* $M = K$, $\mathcal{B} = \mathcal{D}$ [*so that for all* $j \notin \mathcal{B}$, X_j *are i.i.d.* $N(0, \sigma_0^2)$], *and* σ_1 *is fixed. Suppose further that all the assumptions of Theorem* 1 *hold, and there is a* $\delta_+ \in (0,1)$ *such that* (*if* $\mathcal{A}_+$ *is nonempty*)

$$\max_{j \notin \mathcal{A}_+} \|\Sigma^{-1}_{\mathcal{A}_+\mathcal{A}_+} \Sigma_{\mathcal{A}_+ j}\|_1 < \delta_+. \tag{35}$$

Then, given $c_8 \geq 1$, *for all* $\mu_n \geq 0$, *for large enough* n,

$$\mathbb{P}(\widehat{\mathcal{A}}^{\mathrm{LASSO}}_{\mu_n} \neq \mathcal{A}) \geq 1 - n^{-c_8}, \tag{36}$$

where $\widehat{\mathcal{A}}^{\mathrm{LASSO}}_{\mu_n} = \{j \in \mathcal{P} : \widehat{\theta}^{\mathrm{LASSO},\mu_n}_j \neq 0\}$, *where*

$$\widehat{\theta}^{\mathrm{LASSO},\mu_n} = \arg\min_{\zeta \in \mathbb{R}^p} \frac{1}{n}\|Y - \mathbf{X}\zeta\|_2^2 + \mu_n \|\zeta\|_1. \tag{37}$$

Proposition 5 shows that if $\alpha > 1 - 2\kappa_2$, so that $\eta_n = o(n^{-(1-\alpha)/2})$, and the assumptions of Proposition 5 are satisfied, then the SPC + LASSO approach [solving the optimization problem (20) or (19)] can identify $\mathcal{A}$ with appropriate choice of penalization parameter μ_n (as indicated in Theorem 1) while LASSO cannot, with any choice of the penalty parameter.

5. Classification problems and further topics. The preconditioning idea has potential application in any supervised learning problem in which the number of features greatly exceeds the number of observations. A key component is the availability of a consistent estimator for the construction of the preconditioned outcome variable.

For example, preconditioning can be applied to classification problems. Conceptually, we separate the problems of (a) obtaining a good classifier and (b) selecting a small set of good features for classification. Many classifiers, such as the support vector machine, are effective at finding a good separator for the classes. However they are much less effective in distilling these features down into a smaller set of uncorrelated features.

Consider a two-class problem, and suppose we have trained a classifier, yielding estimates $\hat{p}_i$, the probability of class 2 for observation $i = 1, 2, \ldots, N$. Then in the second stage, we apply a selection procedure such as forward stepwise or the LASSO, to an appropriate function of $\hat{p}_i$; the quantity $\log[\hat{p}_i/(1-\hat{p}_i)]$ is a logical choice.

We generated data as in example of Section 3; however, we turned it into a classification problem by defining the outcome class g_i as 1 if $y_i < 0$ and 2 otherwise.

We applied the nearest shrunken centroid (NSC) classifier of Tibshirani et al. [17], a method for classifying microarray samples. We applied forward stepwise regression both to g_i directly (labeled FS), and to the output $\log(\hat{p}_i/(1-\hat{p}_i))$ of the NSC classifier (labeled NSC/FS).

The results of 10 simulations are shown in Figure 5. We see that NSC/FS does not improve the test error of FS, but as shown in the bottom left panel, it does increase the number of "good" predictors that are found. This is a topic of further study.

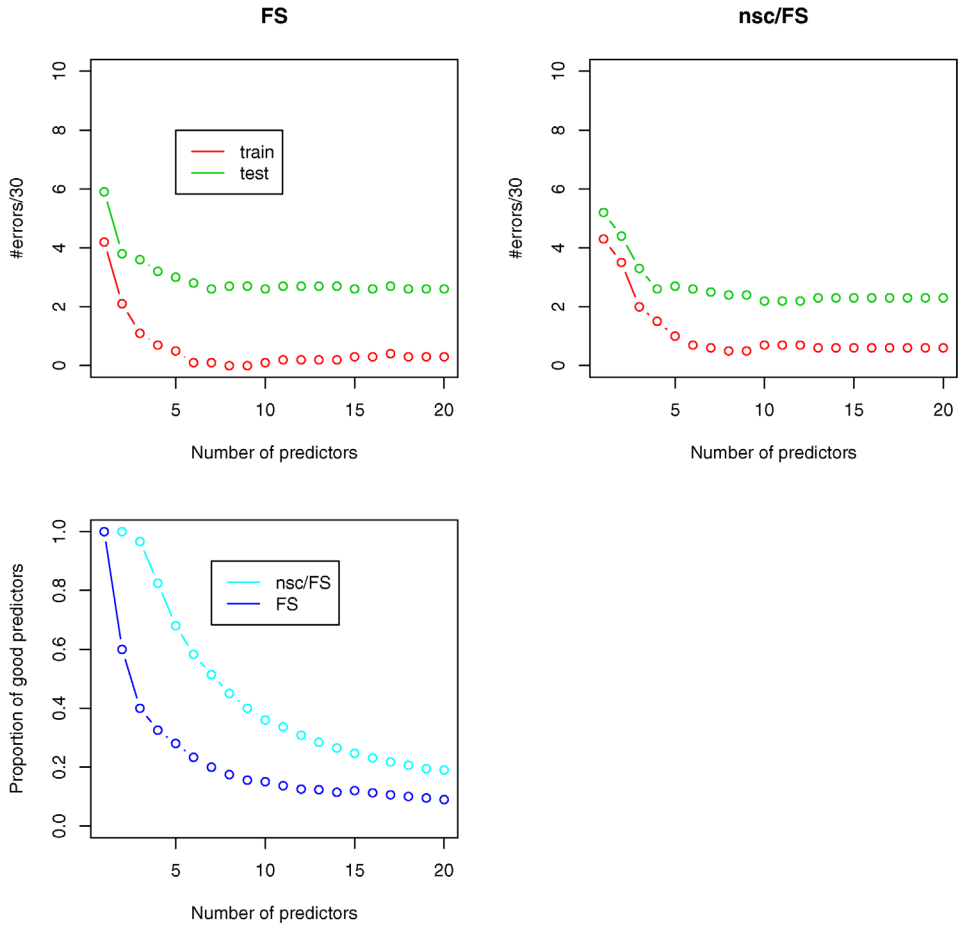

FIG. 5. *Results of applying preconditioning in a classification setting. Top left panel shows the number of test misclassification errors from forward stepwise regression*; *in the top right panel we have applied forward stepwise regression to the preconditioned estimates from nearest shrunken centroid classifier. The proportion of good predictors selected by each method is shown in the bottom left.*

APPENDIX

A full version of this paper that includes the Appendix is available at http://www-stat.stanford.edu/~tibs/ftp/precond.pdf and also in arXiv archive.

Acknowledgments. We thank the referees and editors for comments that led to improvements in this work.

REFERENCES

[1] BAIR, E., HASTIE, T., PAUL, D. and TIBSHIRANI, R. (2006). Prediction by supervised principal components. *J. Amer. Statist. Assoc.* **101** 119–137. MR2252436

[2] BAIR, E. and TIBSHIRANI, R. (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLOS Biology* **2** 511–522.

[3] DONOHO, D. (2004). For most large underdetermined systems of equations, the minimal ℓ^1-norm solution is the sparsest solution. Technical report, Stanford Univ.

[4] DONOHO, D. and ELAD, M. (2003). Optimally sparse representation from overcomplete dictionaries via ℓ^1-norm minimization. *Proc. Natl. Acad. Sci. USA* **100** 2197–2202. MR1963681

[5] EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression (with discussion). *Ann. Statist.* **32** 407–499. MR2060166

[6] FAN, J. and LI, R. (2005). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

[7] FAN, J. and PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32** 928–961. MR2065194

[8] KALBFLEISCH, J. and PRENTICE, R. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York. MR0570114

[9] KNIGHT, K. and FU, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28** 1356–1378. MR1805787

[10] MEINSHAUSEN, M. (2005). Lasso with relaxation. Research Report 129, ETH Zürich.

[11] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

[12] OSBORNE, M., PRESNELL, B. and TURLACH, B. (2000). On the lasso and its dual. *J. Comput. Graph. Statist.* **9** 319–337. MR1822089

[13] PARK, M. Y. and HASTIE, T. (2006). An l1 regularization-path algorithm for generalized linear models. Unpublished manuscript.

[14] PAUL, D. (2005). Nonparametric estimation of principal components. Ph.D. dissertation, Dept. Statistics, Stanford Univ.

[15] SHEN, X. and YE, J. (2002). Adaptive model selection. *J. Amer. Statist. Assoc.* **97** 210–221. MR1947281

[16] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[17] TIBSHIRANI, R., HASTIE, T., NARASIMHAN, B. and CHU, G. (2001). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA* **99** 6567–6572.

[18] ZHAO, H., TIBSHIRANI, R. and BROOKS, J. (2005). Gene expression profiling predicts survival in conventional renal cell carcinoma. *PloS. Med.* **3**(1) e13.

[19] ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

[20] ZOU, H. (2005). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

D. PAUL
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, DAVIS
DAVIS, CALIFORNIA 95616
USA
E-MAIL: debashis@wald.ucdavis.edu

E. BAIR
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305
USA
E-MAIL: ebair@stat.stanford.edu

T. HASTIE
R. TIBSHIRANI
DEPARTMENTS OF STATISTICS AND HEALTH,
RESEARCH AND POLICY
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305
USA
E-MAIL: hastie@stat.stanford.edu
tibs@stat.stanford.edu