Biostatistics (2016), **17**, 2, *pp.* 364–376
doi:10.1093/biostatistics/kxv049
Advance Access publication on November 27, 2015

Sparse regression and marginal testing using cluster prototypes

STEPHEN REID*

Department of Statistics, Stanford University, 390 Serra Mall, Stanford, CA 94305, USA
sreid@stanford.edu

ROBERT TIBSHIRANI

Department of Statistics, Stanford University, 390 Serra Mall, Stanford, CA 94305, USA and Department of Health Research and Policy, Stanford University, 150 Governor's Lane, Stanford, CA 94305, USA

Summary

We propose a new approach for sparse regression and marginal testing, for data with correlated features. Our procedure first clusters the features, and then chooses as the cluster prototype the most informative feature in that cluster. Then we apply either sparse regression (lasso) or marginal significance testing to these prototypes. While this kind of strategy is not entirely new, a key feature of our proposal is its use of the post-selection inference theory of Taylor *and others* (2014, Exact post-selection inference for forward stepwise and least angle regression, Preprint, arXiv:1401.3889) and Lee *and others* (2014, Exact post-selection inference with the lasso, Preprint, arXiv:1311.6238v5) to compute *exact* p-values and confidence intervals that properly account for the selection of prototypes. We also apply the recent "knockoff" idea of Barber and Candès (2014, Controlling the false discovery rate via knockoffs, Preprint, arXiv:1404.5609) to provide exact finite sample control of the FDR of our regression procedure. We illustrate our proposals on both real and simulated data.

Keywords: Clustering; Correlated predictors; Knockoff; Lasso; Marginal screening; Post-selection inference.

1. Introduction

We consider the linear model setup:

$$y = X\beta + \epsilon, \tag{1.1}$$

where y is the $n \times 1$ response vector, X an $n \times p$ matrix of predictors, β a $p \times 1$ vector of true regression coefficients and ϵ an $n \times 1$ vector of errors—usually assumed to have n-dimensional Gaussian distribution $N(0, \sigma^2 I)$. Although not critical for the ideas of this paper, we are especially interested in the $p > n$ case. Since our matrix is not of full column rank, least squares regression cannot be used for further analysis of the problem and we proceed with a regression method that comprises of both variable selection and

*To whom correspondence should be addressed.

parameter estimation. Examples include the lasso of Tibshirani (1996), the elastic net of Zou and Hastie (2005), principal component and forward and backward stepwise regression. We focus primarily on the lasso.

Furthermore, suppose that the columns of X share substantial empirical correlation—artifacts of a hypothetical underlying covariance matrix for the columns that exhibits significant block structure. In this instance, the lasso essentially selects randomly among a set of highly correlated signal variables, entering only one of them, and it struggles to recover the true set of variables. Another, largely unaddressed, issue is that of a lack of interpretability. The lasso merely provides a set of predictive variables. It is silent on the grouping (correlation) structure amongst the variables or indeed the estimated effect on the response of those unselected variables. Of course, this tool is not designed for such a purpose.

Our goal in this paper is to provide the owner of a dataset, exhibiting significant correlation amongst its columns, with a procedure to discover the column groupings and a single representative prototype from each group. Subsequent sparse regression or marginal testing is then performed on these prototypes. Critically, our method allows us to leverage the powerful results of Lee *and others* (2014), Taylor *and others* (2014), and Lee and Taylor (2014) to provide *valid* confidence intervals and p-values (testing for nullity) of the effect sizes of these prototypes. This is true even after their selection as prototypes *and* their participation in subsequent sparse regression. Examples of "effects" on which inference could be done are the components of the vector $\beta_M = (X_M^\top X_M)^{-1} X_M^\top \mu$, where X_M is the matrix X reduced to the columns in the set, M, selected by some variable selection procedure (like the lasso), and β_M contains the population regression coefficients of a subsequent regression only on the columns selected to M. Inference will be performed conditional on the selection event that fruited M.

An additional feature: we can use the group structure learned in the first step and generate similar confidence intervals and p-values for those variables not selected, but correlated with, the prototypes. Note that this provides a highly interpretable snapshot of the data: a grouping structure, effect size estimates of most predictive prototypes of the grouping structure as well as for the also-rans in each of these groups. Our proposal comprises of a series of distinct steps:

(1) *Grouping* the columns of the predictor matrix X (assumed fixed), using either pre-defined groups from the problem context or a clustering method.
(2) *Prototype extraction*, one from each cluster, by screening on the marginal correlation each cluster member has with the response y.
(3) *Subsequent regression analysis* on the selected prototypes, here specifically the lasso, what we call the `Protolasso`, or *marginal testing* of the prototypes, what we call `Prototest`. We use the theory of *post-selection inference* to obtain *exact* p-values and confidence intervals that properly account for the selection at every stage.

As an illustration of the interpretive richness of our method, consider Figure 1. The caption describes how data were generated. A single pair of X and y was generated. To this data we fit the ordinary lasso, used cross-validation to select the optimal number of variables and used the post-selection inference tools of Lee *and others* (2014) to generate confidence intervals for the selected variables. We also subjected the data to our `protolasso` procedure. We describe this procedure in later sections.

This paper is organized as follows. In Section 2, we review some related work in the literature. Section 3 describes the clustering of the features and prototype extraction. In Section 4, we review the theory of post-selection inference and give details and examples of our `Protolasso` proposal. The `Prototest` of Section 5 carries out marginal testing of the selected prototypes. We discuss FDR control for `protolasso` via knockoffs in Section 6. Appendix E of supplementary material available at *Biostatistics* online gives details of the proposed gap statistic for choosing the number of clusters. We conclude with discussion in Section 7.

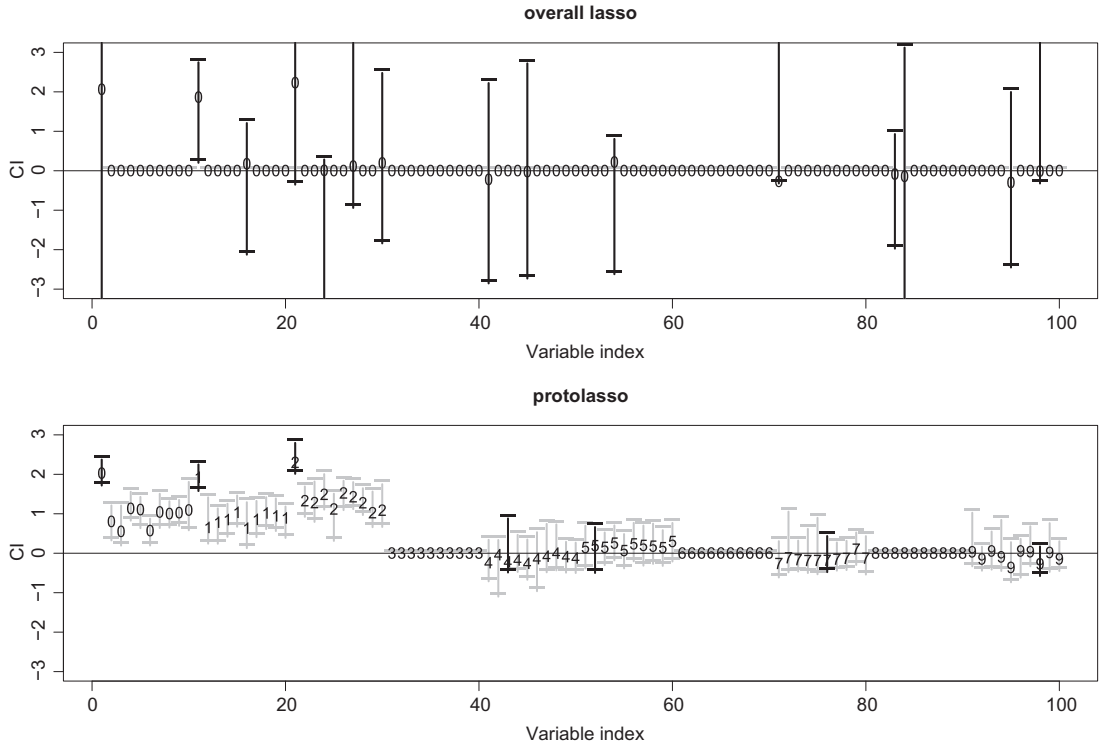

Fig. 1. Confidence interval plots for selected variables from ordinary lasso (*top panel*) and `protolasso` (*bottom panel*). Predictor matrix X has $n = 50$ and $p = 100$ columns, in groups of size 10. Predictors in each group share pairwise correlations of size $\rho = 0.5$, but are uncorrelated with all other predictors. Digit characters represent the true target of the confidence interval (the partial correlation of the appropriate predictor with the response, given the selected model), and also represent the grouping detected by `protolasso`. Confidence intervals are represented by vertical bars and horizontal endpoints: black for selected predictors/prototypes; gray for those obtained by swapping out the cluster prototype for another predictor in the cluster. If predictors/clusters prototypes were not selected, we do not construct a confidence interval for them. Original β vector had indices 1, 11, and 21 non-zero, with value set to 2. Notice that the ordinary lasso detects no group structure (all digits are 0) and seems to construct relatively wide intervals for selected variables (possibly due to the correlations amongst them). Protolasso, on the other hand, provides a group structure, seemingly narrower intervals for the selected prototypes *and* the ability to construct CIs for non-prototype variables in each cluster, behaving as if they were the prototype.

2. RELATED WORK

The lasso has become a widely used tool for linear regression with simultaneous variable selection. There has been much theoretical development of the method's variable selection and screening properties, under a variety of assumptions. References include Meinhausen and Buhlmann (2006), Zhao and Yu (2006), Van de Geer (2007), Zhang and Huang (2008), Van de Geer and Buhlmann (2009), Meinhausen and Yu (2009), Bickel *and others* (2009), and Sun and Zhang (2011).

These positive findings are heartening. However, it has been noted that the lasso performs poorly in the presence of predictors with high empirical correlation. Other methods like the *elastic net* of Zou and Hastie (2005), *OSCAR* of Bondell and Reich (2008), and the *clustered lasso* of She (2010) have been developed to address this shortcoming. These consider different penalty terms in order to encourage more acceptable behavior in the face of highly correlated predictors.

Recently, Campbell and Allen (2015) have studied the Exclusive Lasso. It uses a combination of ℓ_1 penalization within clusters and ℓ_2 penalization between clusters to generate at least one (often exactly one) non-zero coefficient per cluster. Although the method shows promising predictive performance, we do not have closed form distributional results to facilitate the seamless inference that we can perform with our methods.

Another set of methods cluster the variables first and then fit a model. Some methods do these two steps sequentially (clustering first and then fitting a model to some cluster representatives); others simultaneously. Examples of the former include principal component regression, *gene shaving* of Hastie *and others* (2000), *tree harvesting* of Hastie *and others* (2001), averaged gene expressions of Park *and others* (2007) and the canonical correlation clustering and subsequent sparse regression of Buhlmann *and others* (2013). The method of Dettling and Buhlmann (2004) and *OSCAR* represent the latter.

The above methods all attempt to find clusters of variables and combine variables within a cluster to support good predictive ability in the subsequent (or concomitant) model fit. Our method also proceeds sequentially by first clustering variables into highly correlated groups. A second step differentiates our method from the others: once the clusters have been found, we extract a single representative prototype from the cluster membership. We do not combine all members of the cluster by averaging or projecting onto principal component directions. The prototypes are chosen so as to preserve subsequent predictive ability as much as possible.

Once we have constructed clusters and extracted prototypes, the user can proceed as they wish. We, however, follow Buhlmann *and others* (2013) and perform a sparse regression after prototype selection. The novelty of our method (and the source of its rich interpretability) is our liberal use of the post selection inference framework of Taylor *and others* (2014) and Lee *and others* (2014). We first discuss the clustering algorithm and then proceed to post selection inference results.

3. Clustering and prototype extraction

The first step in our procedure is the estimation of the grouping structure of the features. In some problems, these groupings may be pre-defined, for example, gene sets in microarray studies, organizing genes into functional units. In this case, our procedure will make use of these groups.

In many cases, however, no *a priori* grouping is available for the features and unsupervised clustering tools can be applied. Fortunately, the post selection results of Lee *and others* (2014) (the backbone of the ultimate interpretability of our method) all go through, conditional on the predictor matrix X. Any clustering performed on the columns of X need not be taken into account when constructing these inferences later on, provided the clustering is truly unsupervised and receives no input from the response y.

As such, we do not prescribe to the user any particular clustering method. We are free to use k-means, k-medoids, or any of the hierarchical clusterings: complete, single, and minimax linkage (Bien and Tibshirani, 2011), for example. We proceed with hierarchical clustering methods. Minimax clustering has the added advantage of returning a clustering of the variables and a prototype for each cluster. These prototypes are determined solely on merit of X, with no input from y and so may lack some predictive power. They could be used immediately in further analysis, but we propose a different set of prototypes that can be extracted from the output of any clustering method. They are described in a subsequent paragraph.

We propose a gap statistic procedure very similar to that of Tibshirani *and others* (2001) for automatically selecting the number of clusters. To maintain the flow of the exposition, we defer details of this procedure to Appendix E of supplementary material available at *Biostatistics* online.

3.1 *Prototype extraction*

Having identified clusters, the next step is to extract prototypes—one from each cluster. By "prototype", we mean a single cluster representative, chosen from *among the members* of the cluster. We exclude the

possibility of using a point wise mean or first principal component of the points in the cluster. These have already been considered by the likes of Buhlmann *and others* (2013). Furthermore, we wish to select prototypes with some recourse to the response. Perhaps this imbues subsequent regressions and tests with more power. We must be careful to allow for this supervised prototype selection in subsequent inference though. Finally, a single representative of a cluster enhances interpretability, as the investigator now has a single entry point for further examination of a selected cluster.

Once we make use of y in our prototype selection, and proceed with these prototypes in further analysis, any inference downstream must take this selection effect into account. Luckily, we can leverage the post selection inference framework of Lee *and others* (2014) to account for this selection effect, provided the prototype selection method is simple enough, i.e. linear in y, after conditioning on minimal additional information. One possibility would be to use the "least squares prototype"—that is, regress y on the features in the cluster and used the fitted values as the "prototype." This would capture the joint signal in all of the cluster members, and would be particularly appropriate when modeling genesets in genomic studies.

However, in this paper, we focus on simple marginal correlation screening. In particular, given a clustering with K clusters, written as $\{C_1, C_2, \ldots, C_K\}$ where $\bigcup_{k=1}^{K} C_k = \{1, 2, \ldots, p\}$ and $C_j \cap C_k = \phi$ for $j \neq k$, we extract prototypes $\hat{P} = \{\hat{P}_1, \hat{P}_2, \ldots, \hat{P}_K\}$ where each $\hat{P}_k \in \{1, 2, \ldots, p\}$ and $\hat{P}_k \in C_k$ such that $\hat{P}_k = \text{argmax}_{j \in C_k} |x_j^\top y|$.

It turns out that the selection of prototypes in this way is easily translated into the framework of Lee *and others* (2014) and hence easily incorporated into subsequent inference. Even at this juncture, it is possible to make inferences (perform significance tests and construct confidence intervals) for these prototypes. Particular examples include inferences on the individual marginal correlations with y and partial correlations if all (or some) of the prototypes are used in a subsequent least squares regression. Details of post selection inference are discussed in the next section.

Figure 2 demonstrates that the variable selection performance of the `protolasso` procedure (clustering and then marginal correlation screening in each cluster) is not much worse than that of two other selection procedures. The first of the other methods is a marginal screening method, each time selecting as many variables as there are clusters in the `protolasso` method, retaining those with the largest absolute marginal correlation with y. The second is the lasso, applied to the entire dataset, with the regularization parameter chosen so as to ensure the same number of variables is selected.

The lasso performs admirably, over all correlations—a function of the sheer size of the signal. The `protolasso` procedure is not far behind, selecting the non-prototypes in the signal clusters far less often than does the other two methods, while selecting some of the noise variables more often. This is a function of selecting a prototype from each cluster, many of which are filled only with noise variables. Selection performance degrades at very high correlations. However, at these correlations it is very difficult to distinguish between signal and non-signal variables in the same cluster anyway.

The reader should note that our method is not designed to be competitive in the selection of signal variables. There are other methods more effective at this endeavor. The merit of our method is its rich interpretability. We show the output of this particular simulation merely to demonstrate that variable selection performance is not too critically compromised.

4. POST SELECTION INFERENCE

Taylor *and others* (2014) and Lee *and others* (2014) present a powerful framework for inference after variable selection that takes the response into account. They consider the usual linear model, as in Equation (1.1), assuming in particular that $\epsilon \sim N(0, \Sigma)$. Let $\mu = X\beta$.

Their focus is inference on linear combinations $\eta^\top \mu$ *after* the application of a variable selection technique. Variable selection—a process usually privy to the response y—complicates post fit inference. Given

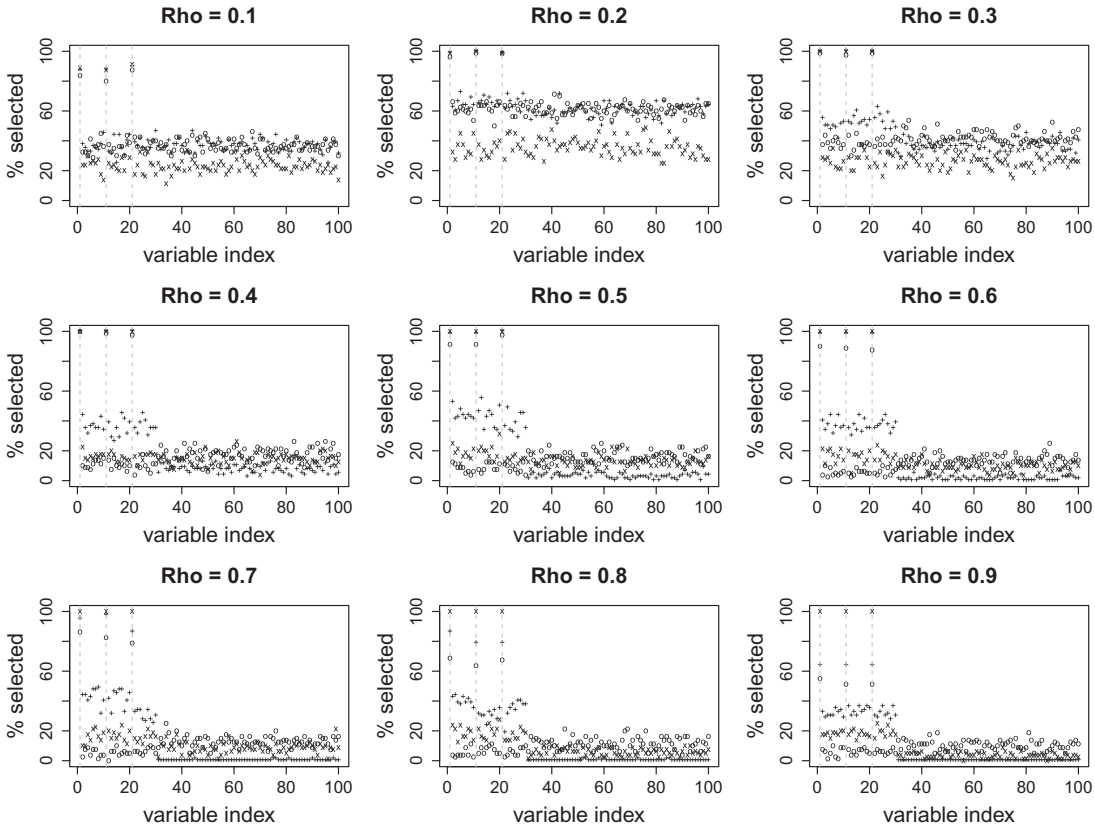

Fig. 2. Dataset has $n = 50$ rows and $p = 100$ columns, divided into 10 groups of 10 columns each. Columns within a group share pairwise correlation ρ, which is set to $\rho = 0.1, 0.2, \ldots, 0.9$ from the top left to the bottom right. Horizontal axes measure the variable number, while vertical axes measure the proportion of times out of $S = 80$ simulation runs in which each variable was selected by each of the methods. Circles (o) represent the `protolasso` procedure, pluses (+) the marginal screening procedure and the crosses (x) the ordinary lasso fit to all the variables. β has 3 non-zero entries—all set to 2—at variables 1, 11, and 21. These are indicated by the three vertical gray dashed lines.

that a set of variables $\hat{M} \subset \{1, 2, \ldots, p\}$ has been selected (often with $|\hat{M}| \leqslant n$), they argue correctly that the usual Gaussian least squares inference on $\beta_{\hat{M}} = X^{+}_{\hat{M}}\mu$ is invalid. Here the subscript denotes the restriction of the columns of X (or the entries of β) to those in the set $\hat{M}$. The superscript $+$ denotes the Moore-Penrose of the matrix in question.

Their main results pertain to the lasso. They show that, for fixed X, and conditioning on both the variables selected by the lasso $\hat{M} = M$ and the signs s_M of the non-zero $\hat{\beta}$, we can construct matrices $A = A^{\text{lasso}}_{M,s}$ and $b = b^{\text{lasso}}_{M,s}$ such that the selection event $\{\hat{M} = M, s_{\hat{M}} = s_M\}$ can be written as a set of *linear* constraints on y: $A^{\text{lasso}}_{M,s_M} y \leqslant b^{\text{lasso}}_{M,s_M}$. Here $s_{M,i} = \text{sign}(\beta_{M,j})$. The authors give explicit formulae for A and b in their paper. The reader is referred there for details. This result can then be used to do inference on $\eta^{\top}\mu$ post selection, by considering the distribution of $\eta^{\top} y | Ay \leqslant b$.

This allows us to characterize the conditional distribution exactly: $F^{[\mathcal{V}^-,\mathcal{V}^+]}_{\eta^{\top}\mu,\eta^{\top}\Sigma\eta}(\eta^{\top} y) | Ay \leqslant b \sim \text{Unif}(0, 1)$, where $F^{[a,b]}_{\mu,\sigma^2}(x) = \frac{\Phi((x-\mu)/\sigma)-\Phi((a-\mu)/\sigma)}{\Phi((b-\mu)/\sigma)-\Phi((a-\mu)/\sigma)}$ – the truncated Gaussian cumulative distribution function (cdf). Again, the exact forms of $\mathcal{V}^-$ and $\mathcal{V}^+$ are given in the paper.

Note that the true utility of this result extends beyond a mere application to the lasso selection procedure. Indeed, any selection procedure that, with perhaps a little extra conditioning information, can be written as a set of linear constraints on y can be subjected to this framework. Other examples of such procedures are marginal correlation screening, as described in Lee and Taylor (2014) and, in the multivariate mean vector estimation literature, the Benjamini-Hochberg procedure and the selection of the largest K elements of y, as described in Reid *and others* (2014). The challenge is to write the selection constraints imposed by a method as a set of linear constraints on y (if possible). This gives one the forms of A and b specific to that method and the rest of the required quantities are easily derived from there.

Our method of clustering, prototyping, and subsequent sparse regression on prototypes is easily cast into this framework. The next subsection describes the selective inference framework for our method. We also subject this method to a simulation study, observing how its performance compares with that of the lasso and marginal screening (both without an initial clustering step). Details of this simulation study are deferred to Appendix A of supplementary material available at *Biostatistics* online.

4.1 *Details of post selection inference for* `Protolasso`

Recall the steps of the `protolasso` and `prototest` procedures given in Section 1, and summarized here for reference. First *group the features*, then *form prototypes* and finally proceed to *model fitting and inference*.

We assume the predictor matrix X fixed. This is also the assumption in the post selection inference framework discussed above. Note that if one assumes X fixed and if the clustering performed on the columns is completely unsupervised (including the method used to estimate the number of clusters), then the post selection distributional results remain unaffected. Identifying the grouping structure in the fixed X has no effect on the linear constraints on y. A new replication of y will *not* change the grouping structure so identified. The first step need not be accounted for in the post selection framework.

The second step does involve y and thus contributes some linear inequalities to the post selection inference. Suppose we are given a cluster C_1, its prototype $\hat{P}_1$ and the sign $\hat{s}_1^{(1)} = \operatorname{sign}(x_{\hat{P}_1}^\top y)$. According to Lee and Taylor (2014), the constraints contributed by selecting the prototype in this cluster can be represented by the matrices

$$A_{\hat{P}_1}^{(1)} = \begin{pmatrix} X_{C_1\setminus\{\hat{P}_1\}}^\top - s_1^{(1)}\mathbf{1}X_{\hat{P}_1}^\top \\ -X_{C_1\setminus\{\hat{P}_1\}}^\top - s_1^{(1)}\mathbf{1}X_{\hat{P}_1}^\top \end{pmatrix}, \quad b_{\hat{P}_1}^{(1)} = \left(\mathbf{0}^\top, \mathbf{0}^\top\right)^\top. \tag{4.1}$$

Collecting the constraints for all the clusters, the entire prototyping step contributes constraint matrices

$$A_{\hat{P}}^{(1)} = \left(A_{\hat{P}_1}^{(1)\top}, A_{\hat{P}_2}^{(2)\top}, \ldots, A_{\hat{P}_K}^{(K)\top}\right)^\top, \qquad b_{\hat{P}}^{(1)} = \left(b_{\hat{P}_1}^{(1)\top}, b_{\hat{P}_2}^{(2)\top}, \ldots, b_{\hat{P}_K}^{(K)\top}\right)^\top, \tag{4.2}$$

where we condition on the set of selected prototypes and the signs of their marginal correlations with y, i.e. the event $\{\hat{P} = P, s^{(1)} = s\}$ where $s^{(1)} = (s_1^{(1)}, s_2^{(2)}, \ldots, s_K^{(K)})$.

Here we limit focus to a single prototype per cluster, for ease of exposition. We avoid having to set another tuning parameter (the number of prototypes to select in each cluster) and gain additional interpretability from selecting a *single* representative prototype from a cluster. We need not, however. Indeed, one could imagine that signal in a cluster is distributed over more than one predictor. The selection of additional prototypes is easily accounted for in the A and b matrices above. See Lee and Taylor (2014).

Finally, the third step proceeds by fitting the lasso on the columns of $X_{\hat{P}}$. Following Lee *and others* (2014), we can construct the constraint matrices for this regression for a fixed regularization parameter λ, once we condition on the set of indices with non-zero coefficients $\hat{M} \subset \hat{P}$ and the signs of these non-zero

coefficients $\hat{s}^{(2)}$. On the conditioning event $\{\hat{P} = P, \hat{M} = M, \hat{s}^{(2)} = s^{(2)}\}$, we have the constraint matrices (where $Q_M = X_M(X_M^\top X_M)^{-1}X_M^\top$):

$$A^{(2)} = \begin{pmatrix} \frac{1}{\lambda} X_{P\setminus M}^\top (I - Q_M) \\ -\frac{1}{\lambda} X_{P\setminus M}^\top (I - Q_M) \\ -\mathrm{diag}(s^{(2)}) \left(X_M^\top X_M\right)^{-1} X_M^\top \end{pmatrix}, \quad b^{(2)} = \begin{pmatrix} \mathbf{1} - X_{P\setminus M}^\top (X^\top)^+ s^{(2)} \\ \mathbf{1} + X_{P\setminus M}^\top (X^\top)^+ s^{(2)} \\ -\lambda \cdot \mathrm{diag}(s^{(2)})(X_M^\top X_M)^{-1} s^{(2)} \end{pmatrix} \tag{4.3}$$

Overall then, if we condition on the event $\{\hat{P} = P, \hat{M} = M, \hat{s}^{(1)} = s^{(1)}, \hat{s}^{(2)} = s^{(2)}\}$, we have the post selection distribution $\eta^\top y \sim F^{[\mathcal{V}^-, \mathcal{V}^+]}_{\eta^\top \mu, \sigma^2 \eta^\top \eta}$, where $\mathcal{V}^-$ and $\mathcal{V}^+$ are gleaned from $A = (A^{(1)\top}, A^{(2)\top})^\top$, $b = (b^{(1)\top}, b^{(2)\top})^\top$, $s^{(1)}$, $s^{(2)}$ and η as described in Lee *and others* (2014). Assuming that σ^2 is known, we are free to do inference on $\eta^\top \mu$.

4.2 *Inference for the selected prototypes*

Suppose that we have run the `protolasso` procedure and have selected the set $\hat{M} = M$ of column indices from the original X matrix. Recall that this set is obtained after prototype selection (ensuring that at most one variable from each cluster makes it into M), giving $\hat{P} = P$, followed by a lasso only on the set of selected prototypes. Suppose that $P_k \in M$. The post selection inference framework allows us to use $\eta^\top y$ to make inference on the quantity $\eta^\top \mu$ *on the event* $\{\hat{P} = P, \hat{M} = M, \hat{s}^{(1)} = s^{(1)}, \hat{s}^{(2)} = s^{(2)}\} = \{Ay \leqslant b\}$. Here $\eta = (X_M^\top)^+_{j^*}$, where j^* is the index in M corresponding to selected prototype P_k.

Note that μ is from the original, full model of (1.1)—the true, unknown mean of y. The quantity $\eta^\top \mu$ is the *partial* regression coefficient of the selected prototype P_k when we regress the true population mean onto the final set, M, of selected prototypes only. Furthermore, we do inference *conditional* on the selection event. This is equivalent to reducing the sample space of y. Inference using quantities based on y needs to account for the reduced sample space when forming its reference (null) distributions. This is the purpose of the truncated Gaussian distribution.

Considerations pertaining to which variables were selected as prototypes and the subsequent effect this has on the validity of constructed intervals are mooted by doing inference *after conditioning* on the selected prototypes. Taking this set as fixed, we construct intervals for the regression coefficients of the partial model containing only these prototypes.

P-values for this coefficient $\eta^\top \mu$ (when testing for nullity) are gleaned by evaluating the cumulative distribution function described above. Note that these p-values have exactly the same interpretation as classical p-values (as measurew of how extreme our observed test statistics are relative to their null distributions). However, because they are gleaned from the truncated Gaussian distribution which accounts for the reduced sample space imposed on y by our selection event, they are valid post-selection, unlike their classical counterparts. Confidence intervals are obtained by inverting the same function, solving for $\eta^\top \mu$, given data $\eta^\top y$. Indeed, this is how we obtained the solid confidence intervals in lower panel of Figure 1.

If a cluster's prototype is not selected in the second stage, we ignore all variables of that cluster in subsequent analysis. We are loath to state that all the variables are irrelevant, because the first step prototype procedure might miss some individual signal variables. However, our method is geared toward enhancing interpretability by reducing the variables under consideration in a principled way.

4.3 *Inference for other members of selected clusters*

We need not only limit ourselves to inference on the selected prototypes alone. The post selection inference framework allows us to say something about what happens when we swap out a selected prototype for other members of its cluster, treating it as if it had been the prototype instead.

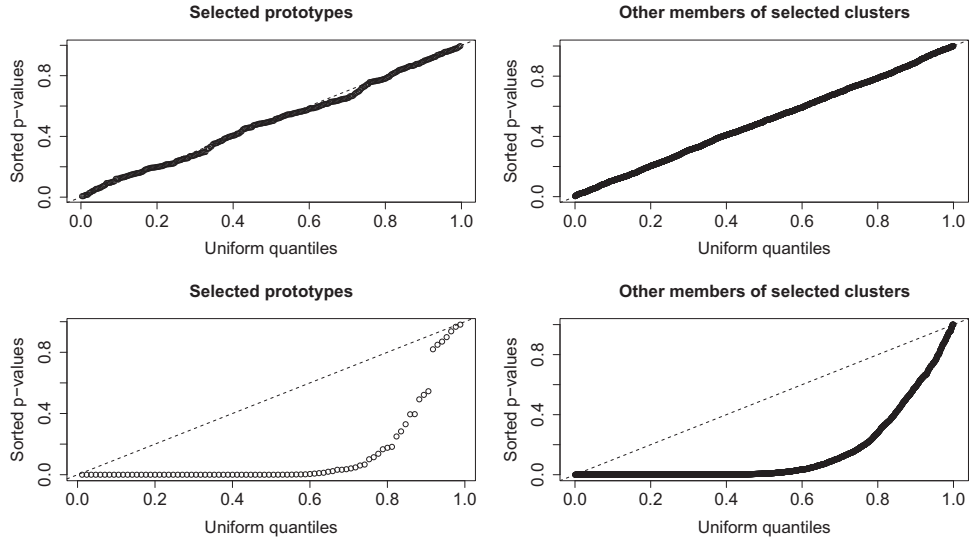

Fig. 3. Dataset as in Figure 1. *Top row*: Sorted p-values (testing for nullity) versus uniform order statistics for selected prototypes (*left*) and other members of selected clusters (*right*) when $\beta = 0$. Notice the close concordance with the 45° line through the origin, as suggested by the theory. *Bottom row:* Same, but for $\beta_1 = \beta_{11} = \beta_{21} = 2$. Notice the sub-uniform behavior of the p-values.

Formally, suppose we have selected prototypes with indices in $\hat{M} = M$ and that one such prototype is $P_k \in M$. Furthermore, suppose that C_k has elements $j_1, j_2, \ldots, j_{|C_k|}$ and we want to exchange $j_1 \neq P_k$ for the the prototype. Let $\eta = (X^{\top}_{M\setminus\{P_k\}\cup\{j_1\}})^{+}_{j}$, where again j is the index in M corresponding to P_k. Again, we make inferences for the *partial* regression coefficient of the swapped-in variable in the partial regression of the true mean μ onto the set of prototypes, without P_k, but including this other member of its cluster. We still have at most one variable from each cluster in this partial regression. Inference still proceeds *conditional on selection*.

We obtained the gray (non-prototype) confidence intervals in the lower panel of Figure 1 in this manner. To convince the reader that these procedures are indeed valid, we generated $S = 100$ replications of this dataset, each time constructing two responses: one with zero signal $\beta = 0$ and one with non-zero signal $\beta_1 = \beta_{11} = \beta_{21} = 2$. We used the gap statistic to select clusters and set the regularization parameter in the second step lasso to ensure that three prototypes were selected. We then computed p-values, under both signal regimes, for the selected prototypes (left panel Figure 3) and all other members of clusters with selected prototypes (right panel Figure 3). Notice that the p-values behave as one would expect—they have uniform distribution for the zero signal ($\beta = 0$) case and a sub-uniform distribution when there is non-zero signal.

5. MARGINAL TESTING: PROTOTEST

In the previous sections, we focussed on the regression problem, applying the lasso to the selected prototypes from clustering of the features, a procedure we called `Protolasso`. In Appendix B of supplementary material available at *Biostatistics* online, we briefly discuss the simpler problem of marginal testing.

6. FDR CONTROL FOR THE PROTOLASSO PROCEDURE

FDR was first introduced in Benjamini and Hochberg (1995) in a multiple testing framework. A subsequent paper of particular interest is that of Barber and Candès (2014), which introduces a variable selection technique guaranteed to control FDR. In a variable selection setting, one would associate R with the number of variables selected and V with the number of null variables selected by the procedure. Their procedure is called *knockoff screening* and proceeds in a sequence of steps: *Formation of the "knockoff" matrix* $\tilde{X}$ of the predictor matrix X. This matrix is constructed so that $X^\top X = \tilde{X}^\top \tilde{X}$ and $X^\top \tilde{X} = X^\top X - \text{diag}(s)$. The authors describe how to construct $\tilde{X}$ and propose how to select the $0 \leqslant s_j \leqslant 1$. *Construction of statistics for each pair of original and knockoff variables.* Each pair gets its own statistic W_j, $j = 1, 2, \ldots, p$. If W_j is large and positive, then there is evidence against the hypothesis that this variable is null. *Definition of data dependent threshold*, $T = \min[t \in \mathcal{W} : \frac{1+\#\{j:W_j \leqslant -t\}}{\#\{j:W_j \geqslant t\} \vee 1} \leqslant q]$, where $\mathcal{W} = \{|W_j| : j = 1, \ldots, p\} \setminus \{0\}$. *Select features* with $W_j \geqslant T$ to control FDR at q.

The shrewdness of the method is in the manner in which they construct the knockoffs and the statistics W_j. They present many examples of potential W_j constructions, but note that their FDR result holds as long as the W_j obey two properties, both detailed in the reference: a sufficiency and a antisymmetry property.

One example of valid W_j, mentioned in the paper, and used by us later in the section, is to fit a lasso regression of y on $[X\ \tilde{X}]$, storing the largest value of the regularization parameter λ such that the variable enters the model, $Z_j = \sup\{\lambda : \hat{\beta}_j(\lambda) \neq 0\}$, with the knockoff equivalent $\tilde{Z}_j$ similarly defined. Setting $W_j = Z_j - \tilde{Z}_j$ achieves the desired properties.

Proof of their results follows from a super-martingale argument which flows quite elegantly, considering the obviously complicated distributional properties of the W_j and T. In fact, the crux of the result hinges on a succession of lemmas (numbers 1, 2, and 3 in the reference) that establish certain exchangeability results, making the rest of the analysis essentially independent of the underlying distribution of the W_j.

In the remainder of this section, we describe a method modifying the knockoff procedure to operate at the prototype level, meshing it nicely with the initial clustering-prototyping step championed in this paper. We design the procedure to allow replication of Lemmas 1, 2, and 3 of Barber and Candès (2014), hence establishing FDR control for our procedure at the prototype level. The section concludes with a brief experiment demonstrating an application of our knockoff method to `Protolasso`.

6.1 *A Knockoff procedure for* `protolasso`

Suppose we have column centered and standardized the predictor matrix X. Recognizing that the construction of knockoff $\tilde{X}$ ensures $X^\top X = \tilde{X}^\top \tilde{X}$, one realizes that any clustering based on the correlation metric detailed above produces the *same* clustering on both sets of columns. An initial instinct is to cluster the columns of X (and, by extension, for $\tilde{X}$) and then finding maximal marginal correlation prototypes for each separately, delivering prototypes $P = \{P_1, P_2, \ldots P_K\}$ for the K clusters of the columns of X and similarly $\tilde{P} = \{\tilde{P}_1, \tilde{P}_2, \ldots, \tilde{P}_K\}$ for $\tilde{X}$. Note that the separate computation of the two sets of prototypes does not guarantee that $P_j = \tilde{P}_j$, despite the initial clustering being the same. This is because the maximal marginal correlation column (with response y) in a given cluster need not be the same in X as it is in $\tilde{X}$. The knockoff procedure alters correlations with the response. One would then venture to proceed by forming matrices X_P and $\tilde{X}_{\tilde{P}}$ and performing the knockoff procedure as described by Barber and Candès (2014) on the response y with augmented matrix $[X_P\ \tilde{X}_{\tilde{P}}]$. Although we believe this procedure *could* control FDR, we found in experiments that it has very low power (ability to detect prototypes in clusters containing signal variables). We suspect that this has to do with the separate selection of prototypes in the original and knockoff matrices, especially in matrices where we encounter high correlation. In high correlations, prototypes are less likely to be the actual signal variables from a cluster (since every variable is a good surrogate for all the others and non-signal variables may present as having high correlation with the

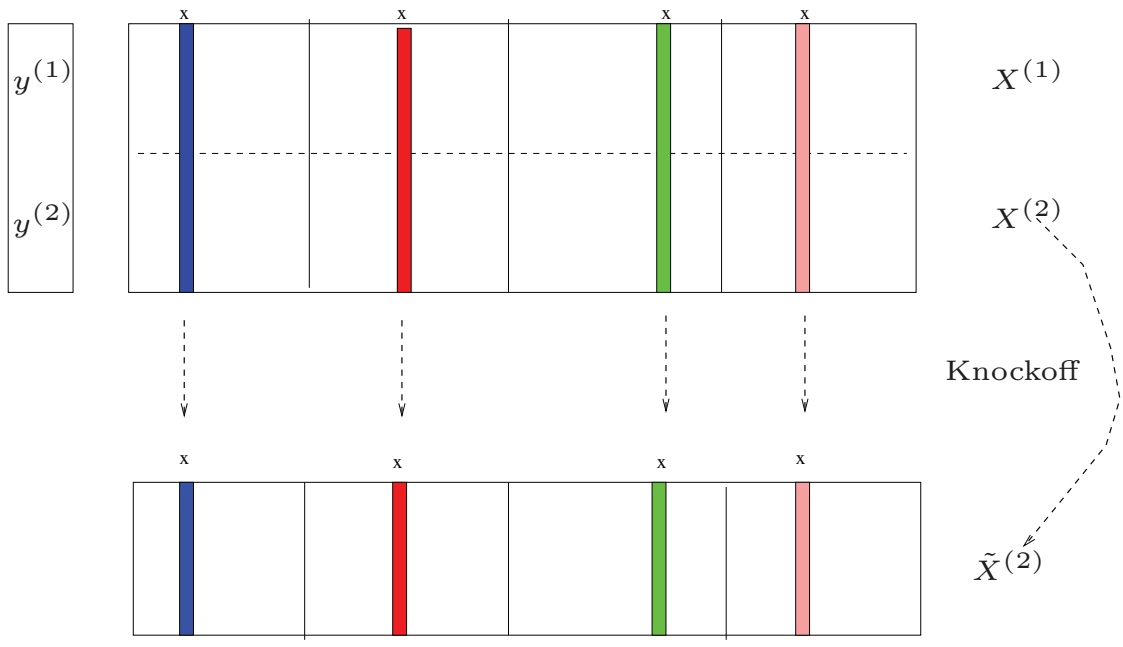

Fig. 4. Schematic of our knockoff strategy with sample splitting: Columns of original matrix are clustered using all of X. Solid vertical black lines delimit clusters. There are four clusters in the figure. Notice how the clusters are translated exactly to the knockoff matrix. Original matrix is then split into $X^{(1)}$ and $X^{(2)}$ and response into $y^{(1)}$ and $y^{(2)}$ (horizontal dotted line in top part of the figure). Prototypes are found for the clusters using only $X^{(1)}$ and $y^{(1)}$. These prototypes are highlighted by shaded vertical bars. Notice how they too are translated directly over to the knockoff matrix. Knockoff matrix $\tilde{X}^{(2)}$ is obtained from the *whole* $X^{(2)}$ before reducing to prototypes. Notice that a knockoff of $X^{(1)}$ is not created.

response). Also, we choose the most correlated knockoff variable from the knockoff cluster, which reduces the apparent strength of the original variable in the subsequent computation of W_j (even if this original variable comes from a signal cluster). We see fewer large positive W_j and knockoff screening tends not to select any variables. In sum: prototypes have sight of y before subsequent analysis and selected prototypes tend to mismatch over original and prototype matrices: $P_j \neq \tilde{P}_j$. This reduces power.

Our procedure addresses each of these shortcomings and orders steps in such a way to ensure the exchangeability lemmas of Barber and Candès (2014) still hold. The procedure is illustrated in Figure 4 and detailed as follows: *Cluster columns* of X and select number of clusters K, producing clustering $\{C_1, C_2, \ldots, C_K\}$. *Split the data* by rows into two (roughly) equal parts: $y = \binom{y^{(1)}}{y^{(2)}}$ and $X = \binom{X^{(1)}}{X^{(2)}}$. *Prototype clusters* via maximal marginal correlations, as before, using only $y^{(1)}$ and $X^{(1)}$. This yields prototype set $\hat{P} = \{\hat{P}_1, \hat{P}_2, \ldots, \hat{P}_K\}$. $y^{(1)}$ and $X^{(1)}$ are now excluded from further analysis. *Form knockoff matrix* $\tilde{X}^{(2)}$ from $X^{(2)}$ as described in Barber and Candès (2014). It is essential that we form the knockoff matrix here using all columns of $X^{(2)}$. This is to ensure that the exchangeability results hold for our procedure. Details are in Appendix F. *Reduce to matrices* $X^{(2)}_{\hat{P}}$ and $\tilde{X}^{(2)}_{\hat{P}}$ by selecting the relevant (and same) columns in $X^{(2)}$ and $\tilde{X}^{(2)}$, respectively. Then *proceed with knockoff screening* using $y^{(2)}$ and $[X^{(2)}_{\hat{P}}\ \tilde{X}^{(2)}_{\hat{P}}]$.

We show in the Appendix F of supplementary material available at *Biostatistics* online how this procedure replicates the exchangeability lemmas in Barber and Candès (2014). In Appendix C of supplementary

material available at *Biostatistics* online, we perform an experiment to study the power of the procedure. Our procedure exactly controls FDR in finite samples, as summarized in the following lemma:

LEMMA 6.1 For any $q \in [0, 1]$, let $W_j^{(2)}$, $j \in \hat{P}$ be the statistics testing the knockoff-original pairs in the knockoff procedure for `protolasso`, where $\hat{P} = \{\hat{P}_1, \hat{P}_2, \ldots, \hat{P}_K\}$ is the set of prototypes selected in the first stage of the `protolasso` procedure. Furthermore, let $T = \min[t \in \mathcal{W} : \frac{1+\#\{j \in \hat{P} : W_j^{(2)} \leqslant -t\}}{\#\{j \in \hat{P} : W_j^{(2)} \geqslant t\} \vee 1} \leqslant q]$, where $\mathcal{W} = \{|W_j^{(2)}| : j \in \hat{P}\} \setminus \{0\}$. Finally, let $\hat{S}_{\hat{P}} = \{j \in \hat{P} : W_j^{(2)} \geqslant T\}$. Then, $E[\frac{\#\{j : \beta_{\hat{P}_j} = 0 \text{ and } j \in \hat{S}_{\hat{P}}\}}{\#\{j : j \in \hat{S}_{\hat{P}}\} \vee 1}] \leqslant q$, where the expectation is taken over the distribution of $y^{(1)}$, which generates the prototype set $\hat{P}$, and that of $y^{(2)}$, which generates the statistics $W_j^{(2)}$.

7. DISCUSSION

We have introduced a coherent procedure for clustering, prototyping, and subsequent analysis of datasets with groups of correlated variables. The biggest selling point of our procedure is our use of the post-selection framework of Lee *and others* (2014) to obtain *exact* p-values in subsequent (i.e. post prototyping) regression and/or testing. Furthermore, we show how the recently proposed knockoff procedure of Barber and Candès (2014) can be adapted to our `protolasso` procedure, guaranteeing FDR control with reasonable power to detect variables in true signal clusters. Subsequent research will include further development and study of the `prototest` procedure, as marginal testing of correlated features is an important and challenging problem, for example in computational biology.

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

We thank Peter Buhlmann and Emmanuel Candes for helpful conversations. Robert Tibshirani was supported by NSF grant DMS-9971405 and NIH grant N01-HV-28183. *Conflict of Interest*: None declared.

FUNDING

R.T. was supported by the NSF grant DMS-9971405 and NIH grant N01-HV-28183.

REFERENCES

BARBER, R. F. AND CANDÈS, E. (2014). Controlling the false discovery rate via knockoffs. Preprint, arXiv:1404.5609.

BENJAMINI, Y. AND HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**(1), 289–300.

BICKEL, P., RITOV, Y. AND TSYBAKOV, A. (2009). Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics* **37**, 1705–1732.

BIEN, J. AND TIBSHIRANI, R. (2011). Hierarchical clustering with prototypes via minimax linkage. *Journal of the American Statistical Association* **106**(495), 1075–1084.

BONDELL, H. AND REICH, B. (2008). Simultaneous regression shrinkage, variable selection and clustering of predictors with oscar. *Biometrics* **64**, 115–123.

BUHLMANN, P., RUTIMANN, P., VAN DE GEER, S. AND ZHANG, C.-H. (2013). Correlated variables in regression: clustering and sparse estimation. *Journal of Statistical Planning and Inference* **143**, 1835–1871.

CAMPBELL, F. AND ALLEN, G. (2015). Within group variable selection through the exclusive lasso. Preprint, arXiv:1505.07517.

DETTLING, M. AND BUHLMANN, P. (2004). Finding predictive gene groups from microarray data. *Journal of Multivariate Analysis* **90**, 106–131.

HASTIE, T., TIBSHIRANI, R., BOTSTEIN, D. AND BROWN, P. (2001). Supervised harvesting of expression trees. *Genome Biology* **2**, 1–12.

HASTIE, T., TIBSHIRANI, R., EISEN, M., ALIZADEH, A., LEVY, R., STAUDT, L., CHAN, W., BOTSTEIN, D. AND BROWN, P. (2000). 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology* **1**, 1–21.

LEE, J., SUN, D., SUN, Y. AND TAYLOR, J. (2014). Exact post-selection inference with the lasso. Preprint, arXiv:1311.6238v5.

LEE, J. AND TAYLOR, J. (2014). Exact post model selection inference for marginal screening. Preprint, arXiv:1402.5596v2.

MEINHAUSEN, N. AND BUHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Annals of Statistics* **34**, 1436–1462.

MEINHAUSEN, N. AND YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics* **37**, 246–270.

PARK, M., HASTIE, T. AND TIBSHIRANI, R. (2007). Averaged gene expressions for regression. *Biostatistics* **8**(2), 212–227.

REID, S., TAYLOR, J. AND TIBSHIRANI, R. (2014). Post-selection point and interval estimation of signal sizes in gaussian samples. Preprint, arXiv:1405:3340.

SHE, Y. (2010). Sparse regression with exact clustering. *Electronic Journal of Statistics* **4**, 1055–1096.

SUN, T. AND ZHANG, C.-H. (2011). Scaled sparse linear regression. Preprin, arXiv:1104.4595v1.

TAYLOR, J., LOCKHART, R., TIBSHIRANI, R. AND TIBSHIRANI, R. (2014). Exact post-selection inference for forward stepwise and least angle regression. Preprint, arXiv:1401.3889.

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

TIBSHIRANI, R., WALTHER, G. AND HASTIE, T. (2001). Estimating the number of clusters in a dataset via the gap statistic. *Journal of the Royal Statistical Society Series B* **63**(2), 411–423.

VAN DE GEER, S. (2007). The deterministic lasso. *JSM Proceedings*. American Statistical Association.

VAN DE GEER, S. AND BUHLMANN, P. (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics* **3**, 1360–1392.

ZHANG, C.-H. AND HUANG, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *Annals of Statistics* **36**, 1567–1594.

ZHAO, P. AND YU, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research* **7**, 2541–2563.

ZOU, H. AND HASTIE, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B* **67**(2), 301–320.

[*Received April 10, 2015; revised October 19, 2015; accepted for publication October 26, 2015*]