Consistent Group Identification and Variable Selection in Regression with Correlated Predictors

Dhruv B. Sharma, Howard D. Bondell and Hao Helen Zhang*

Last revised on: `May 18, 2010`

Abstract

Statistical procedures for variable selection have become integral elements in any analysis. Successful procedures are characterized by high predictive accuracy, yielding interpretable models while retaining computational efficiency. Penalized methods that perform coefficient shrinkage have been shown to be successful in many cases. Models with correlated predictors are particularly challenging to tackle. We propose a penalization procedure that performs variable selection while clustering groups of predictors automatically. The oracle properties of this procedure including consistency in group identification are also studied. The proposed method compares favorably with existing selection approaches in both prediction accuracy and model discovery, while retaining its computational efficiency.

KEY WORDS: Coefficient shrinkage; Correlation; Group identification; Oracle properties; Penalization; Supervised clustering; Variable selection.

*Dhruv B. Sharma is a graduate student, Howard D. Bondell is an Assistant Professor and Hao Helen Zhang is an Associate Professor in the Department of Statistics at NC State University, Raleigh, NC 27695-8203

1 Introduction

The collection of large quantities of data is becoming increasingly common with advances in technical research. These changes have opened up many new statistical challenges; the availability of high dimensional data often brings with it large quantities of noise and redundant information. Separating the noise from the meaningful signal has led to the development of new statistical techniques for variable selection.

Consider the usual linear regression model setup with n observations and p predictors given by

$$\mathbf{y} = X\beta + \epsilon,$$

where ϵ is a random vector with $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2 I$. Let the response vector be $\mathbf{y}$= $(y_1, ..., y_n)^T$, the j^{th} predictor of the design matrix X be $\mathbf{x}_j$=$(x_{1j}, ..., x_{nj})^T$ for *j=1,...,p* and the vector of coefficients be $\beta = (\beta_1, ..., \beta_p)^T$. Assume the data is centered and the predictors are standardized to have unit L_2 norm, so that $\sum_{i=1}^{n} y_i = 0$, $\sum_{i=1}^{n} x_{ij} = 0$ and $\sum_{i=1}^{n} x_{ij}^2 = 1$ for all *j=1,...,p*. This allows the predictors to be put on a comparative scale, and the intercept may be omitted.

In a high dimensional environment this model could include many predictors that do not contribute to the response or have an indistinguishable effect on the response, making variable selection and the identification of the true underlying model an important and necessary step in the statistical analysis. To this end, consider the true underlying linear regression model structure given by

$$\mathbf{y} = X_o\beta^* + \epsilon,$$

where X_o is the $n \times p_o$ true design matrix obtained by removing unimportant predictors

and combining columns of predictors with indistinguishable coefficients and β^* is the corresponding true coefficient vector of length p_o. For example, if the coefficients of two predictors are truly equal in magnitude, we would combine these two columns of the design matrix by their sum and if a coefficient were truly zero, we would exclude the corresponding predictor. In practice, the discovery of the true model raises two issues; the exclusion of unimportant predictors and the combination of predictors with indistinguishable coefficients. Most existing variable selection approaches can exclude unimportant predictors but fail to combine predictors with indistinguishable coefficients. In this paper we explore a variable selection approach that achieves both goals.

Penalization schemes for regression such as ridge regression (Hoerl and Kennard, 1970) are commonly used in coefficient estimation. Suppose $P_\lambda(\cdot)$ is a penalty on the coefficient vector and λ is its corresponding non-negative penalization parameter. From a loss and penalty setup, estimates from a penalization scheme for the least squares model are given as the minimizers of

$$\|\mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \beta_j\|^2 + P_\lambda(\beta).$$

Penalization techniques for variable selection in regression models have become increasingly popular, as they perform variable selection while simultaneous estimating the coefficients in the model. Examples include nonnegative garrote (Breiman, 1995), least absolute shrinkage and selection operator (LASSO, Tibshirani, 1996), smoothly clipped absolute deviation (SCAD, Fan and Li, 2001), elastic net (Zou and Hastie, 2005), fused LASSO (Tibshirani et al., 2005), adaptive LASSO (Zou, 2006) and group LASSO (Yuan and Lin, 2006). Although these approaches are popular, they fail to combine predictors with indistinguishable coefficients. Recent additions to the literature including octagonal shrinkage and clustering algorithm for regression (OSCAR, Bondell and Reich, 2008) and collapsing and shrinkage in

analysis of variance (CAS-ANOVA, Bondell and Reich, 2009) were studied to address this limitation in the linear regression and ANOVA context respectively.

Supervised clustering is an approach to address the combined effect of predictors with indistinguishable coefficients. The process aims to identify meaningful groups of predictors that form predictive clusters; such as a group of highly correlated predictors that have a unified effect on the response. The OSCAR is a penalization procedure for simultaneous variable selection and supervised clustering. It identifies a predictive cluster by assigning identical coefficients to each element in the group up to a change in sign, while simultaneously eliminating extraneous predictors. The OSCAR estimates can be defined as the solution to

$$\|\mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \beta_j\|^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{1 \le j < k \le p} \max\{|\beta_j|, |\beta_k|\}. \tag{1}$$

The OSCAR penalty contains two penalization parameters λ_1 and λ_2, for $\lambda_1, \lambda_2 \geq 0$. The first, λ_1, penalizes the L_1 norm of the coefficients and encourages sparseness. The second, λ_2, penalizes the pair-wise L_∞ norm of the coefficients encouraging equality of coefficients.

Although the OSCAR performs effectively in practice, we note that the procedure has certain limitations. The method defined in (1) is implemented via quadratic programming of order p^2, which becomes increasingly expensive as p increases. Another limitation of the OSCAR is that it is not an oracle procedure. An oracle procedure (Fan and Li, 2001) is one that should consistently identify the correct model and achieve the optimal estimation accuracy. That is, asymptotically, the procedure performs as well as performing standard least-squares analysis on the correct model, were it known beforehand.

Adaptive weighting is a successful technique to constructing oracle procedures. Zou (2006) showed oracle properties for the adaptive LASSO in linear models and generalized linear models (GLMs) by incorporating data dependent in the penalty. Oracle properties for adaptive LASSO was separately studied in other contexts including survival models by

Zhang and Lu (2007) and least absolute deviation (LAD) models by Wang et al. (2007). The reasoning behind these weights is to ensure that estimates of larger coefficients are penalized less while those that are truly zero have unbounded penalization. Note that all of these procedures define an oracle procedure solely based on selecting variables, not the full selection and grouping structure. Furthermore, an oracle procedure for grouping must also consistently identify the the group of indistinguishable coefficients.

Bondell and Reich (2009) showed oracle properties for the full selection and grouping structure of the CAS-ANOVA procedure in the ANOVA context, using similar arguments of adaptive weighting. In this paper, we show that the OSCAR penalty does not lend itself intuitively to data adaptive weighting. Weighting the pairwise L_∞ norm by an initial estimate, as is the typical case, fails in the following simple scenario. Suppose two coefficients in the pair-wise term are small but should be grouped, then an initial estimate of this quantity would shrink both to zero instead of setting them as equal.

These limitations are the main motivations of this paper. Our goal in this paper is to find an oracle procedure for simultaneous group identification and variable selection, and to address the limitations of existing methods, including the computational burden. The remainder of this paper proceeds as follows. Section 2 proposes a penalization procedure for simultaneous grouping and variable selection along with an algorithm efficient for computation. Section 3 studies theoretical properties of the procedure. Section 4 discusses extensions of the method to GLMs. Section 5 contains a simulation study and the analysis of real examples. A discussion concludes in Section 6.

2 Pairwise Absolute Clustering and Sparsity

2.1 Methodology

In this section, we consider a new penalization scheme for simultaneous group identification and variable selection. We introduce an alternative to the pairwise L_∞ norm in Bondell and Reich (2008) for setting coefficients as equal in magnitude. Note that coefficients with opposite signs are desired to be grouped together in the presence of high negative correlation, leaving the problem equivalent to a sign change of the predictors.

In our setup, the equality of coefficients is achieved by penalizing the pairwise differences and pairwise sums of coefficients. In particular, we propose a penalization scheme with non-negative weights, $\boldsymbol{w}$, whose estimates are the minimizers of

$$\|\mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \beta_j\|^2 + \lambda \{ \sum_{j=1}^{p} w_j |\beta_j| + \sum_{1 \le j < k \le p} w_{jk(-)} |\beta_k - \beta_j| \ + \sum_{1 \le j < k \le p} w_{jk(+)} |\beta_j + \beta_k| \}. \quad (2)$$

The penalty in (2) consists of a weighted L_1 norm of the coefficients that encourages sparseness and a penalty on the differences and sums of pairs of coefficients that encourages equality of coefficients. The weighted penalty on the differences of pairs of coefficients encourages coefficients with the same sign to be set as equal, while the weighted penalty on the sums of pairs of coefficients encourages coefficients with opposite sign to be set as equal in magnitude. The weights are pre-specified non-negative numbers, and their choices will be studied in Section 2.2. We call this penalty Pairwise Absolute Clustering and Sparsity (PACS) penalty and for the remainder of this article we will refer to the PACS procedure in the form given in (2). The PACS objective function is a convex function since it is a sum of convex functions; in particular, if $X^T X$ is full rank then it is strictly convex.

We note here that a specific case of the PACS turns out to be an equivalent repre-

sentation for the OSCAR. It can be shown that

$$\max\{|\beta_j|, |\beta_k|\} = \frac{1}{2}\{|\beta_k - \beta_j| \ + |\beta_j + \beta_k|\}.$$

Suppose $0 \leq c \leq 1$, then the OSCAR estimates can be equivalently expressed as the minimizers of

$$\|\mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \beta_j\|^2 + \lambda\{\sum_{j=1}^{p} c|\beta_j| + \sum_{1 \leq j < k \leq p} 0.5(1-c)|\beta_j - \beta_k| \ + \sum_{1 \leq j < k \leq p} 0.5(1-c)|\beta_j + \beta_k|\}.$$

Hence the OSCAR can be regarded as a special case of the PACS.

The PACS formulation enjoys certain advantages over the original formulation of the OSCAR as given in (1). Like the OSCAR it can be computed via quadratic programming. However, it can also be computed using a local quadratic approximation of the penalty, which is not directly applicable to the original formulation of the OSCAR. The latter strategy is superior to quadratic programming in that quadratic programming becomes expensive in computation and is not feasible for large and even moderate numbers of parameters, while the latter strategy continues to be feasible in these cases, and can be conveniently implemented in standard software.

Furthermore the choice of weighting scheme in the penalty offers the possibility of subjectivity. With a proper reformulation, we show later that it needs only one tuning parameter. The OSCAR has two tuning parameters and this reduction in tuning parameters vastly improves on the cost of computation. This flexibility in choice of weights allows us to explore data adaptive weighting. This is explored in Section 3 where we show that a data adaptive PACS has the oracle property for variable selection and group identification.

Figure 1 goes here.

Figure 1 is an illustration of the flexibility of the PACS approach over the OSCAR

approach in terms of the constraint region in the (β_1, β_2) plane. In figure 1 (a), we see that the OSCAR constraint region has the same shape in all four quadrants. Note that although the shape varies with c in $0 \leq c \leq 1$, it always remains symmetric across the four axes of symmetry. Two particular PACS constraint regions formed with two different choices of weights are seen in figures 1 (b) and 1 (c). As we see from these figures, they have very different shapes and suggest the flexibility of the PACS approach over the OSCAR approach.

2.2 Choosing the Weights

In this section we study different strategies for choosing the weights. The choice of weights offers the possibility of subjectivity which come in various forms. Three choices will be examined in detail: weights determined by a predictor scaling scheme, data adaptive weights for oracle properties, and an approach to incorporate variable correlation into the weights.

2.2.1 Scaling of the PACS Penalty

The weights for the PACS could be determined via standardization. For any penalization scheme, it is important that the predictors are on the same scale so that penalization is done equally. In penalized regression this is done by standardization, for example, each of the columns of the design matrix has unit L_2 norm. In a penalization scheme such as the LASSO, upon standardization, each column of the design matrix contributes equally to the penalty. When the penalty is based on a norm, rescaling a column of the design matrix is equivalent to weighting coefficients by the inverse of this scaling factor (Bondell and Reich, 2009). We will now determine the weights in (2) via a standardization scheme.

Standardization for the PACS is not a trivial matter since it must incorporate the pairwise differences and sums of coefficients. In fact one needs to consider an over-parameterized design matrix that includes the pairwise coefficient differences and sums. Let the vector of

pairwise coefficient differences of length $d = p(p-1)/2$ be given by $\tau = \{\tau_{jk} : 1 \leq j < k \leq p\}$, where $\tau_{jk} = \beta_k - \beta_j$. Similarly, let the vector of pairwise coefficient sums of length d be given by $\gamma = \{\gamma_{jk} : 1 \leq j < k \leq p\}$, where $\gamma_{jk} = \beta_k + \beta_j$. Let $\theta = [\beta^T \ \tau^T \ \gamma^T]^T$ be the coefficient vector of length $q = p^2$ for this over-parameterized model. We have $\theta = M\beta$, where M is a matrix of dimension $q \times p$ given by $\mathrm{M} = [I_p \ D_{(-)}^T \ D_{(+)}^T]^T$, with $D_{(-)}$ being a $d \times p$ matrix of ± 1 that creates η from β and $D_{(+)}$ being a $d \times p$ matrix of $+1$ that creates γ from β. The corresponding design matrix for this over-parameterized design is an $n \times q$ matrix such that $Z\theta = X\beta$ for all β, i.e. $ZM = X$.

Note that Z is not uniquely defined; possible choices include $Z = [X \ 0_{n \times 2d}]$ and $Z = XM^*$, where M^* is any left inverse of M. In particular, choose $Z = XM^-$, where M^- denotes the Moore-Penrose generalized inverse of M.

Proposition 1. *The Moore-Penrose generalized inverse of M is* $M^- = \frac{1}{(2p-1)}[I_p \ D_{(-)}^T \ D_{(+)}^T]$.

Proof of Proposition 1 is given in the appendix. The above proposition allows the determination of the weights via the standardization in the over-parameterized design space. The resulting matrix Z determined using the Moore-Penrose generalized inverse of M is an appropriate design matrix for the over-parametrization. We propose to use the L_2 norm of the corresponding column of Z for standardization as the weights in (2). In particular, these weights are $w_j = 1$, $w_{jk(-)} = \sqrt{2(1 - r_{jk})}$ and $w_{jk(+)} = \sqrt{2(1 + r_{jk})}$ for $1 \leq j < k \leq p$, where r_{jk} is the correlation between the $(j, k)^{th}$ pair of predictors of the standardized design matrix.

2.2.2 Data Adaptive Weights

PACS with appropriately chosen data adaptive weights will be shown to be an oracle procedure. Suppose $\tilde{\beta}$ is a $\sqrt{n}$-consistent estimator of β, such as the ordinary least squares (OLS) estimates. Adaptive weights for the PACS penalty are given by $w_j = |\tilde{\beta}_j|^{-\alpha}$, $w_{jk(-)} =$

$|\tilde{\beta}_k - \tilde{\beta}_j|^{-\alpha}$ and $w_{jk(+)} = |\tilde{\beta}_k + \tilde{\beta}_j|^{-\alpha}$ for $1 \leq j < k \leq p$ and $\alpha > 0$. Such weights allow for less penalization when the coefficients, their pairwise differences, or their pairwise sums are larger in magnitude and penalized in an unbounded manner when they are truly zero.

We note that for $\alpha = 1$, the adaptive weights belong to a class of scale equivariant weights, as long as the initial estimator $\tilde{\beta}$ is scale equivariant. When the weights are scale equivariant, the resulting solution to the optimization problem is also scale equivariant. In such cases the design matrix does not need to be standardized beforehand, since the resulting solution obtained after transforming back is the same as the solution obtained when the design matrix is not standardized. Due to this simplicity, for the remainder of this paper we set $\alpha = 1$.

In practice, the initial estimates can be obtained using OLS or other shrinkage estimates like ridge regression estimates. In studies with collinear predictors, we notice that using ridge estimates for the adaptive weights perform better than those given by OLS estimates. The choice of ridge estimate controls the collinearity by smoothing the coefficients. Ridge estimates selected by AIC are used for adaptive weights in all applications in this paper. Ridge estimates chosen by AIC provide slight regularization while not over shrinking, as opposed to BIC. Note that, since the original data are standardized, the ridge estimates are equivariant. Any other choice would require using the scaling in Section 2.2.1 along with the adaptive weights.

2.2.3 Incorporating Correlation

The choice of weights is subjective and in particular, one may wish to assist the grouping of predictors based on the correlation between predictors. This approach is explored in Tutz and Ulbricht (2009) in the ridge regression context. To this end, consider incorporating correlation in the weighting scheme scheme such as those given by $w_j = 1$, $w_{jk(-)} = (1 - r_{jk})^{-1}$ and $w_{jk(+)} = (1 + r_{jk})^{-1}$ for $1 \leq j < k \leq p$. Intuitively, these weights more heavily penalize

the differences in coefficients when they are highly positively correlated and heavily penalize their sums when they are highly negatively correlated. Though such weighting will not discourage uncorrelated predictors to have equal coefficients, they will encourage pairs of highly correlated predictors to have equal coefficients. Adaptive weights that incorporate these correlations terms are then given by $w_j = |\tilde{\beta}_j|^{-1}$, $w_{jk(-)} = (1 - r_{jk})^{-1}|\tilde{\beta}_k - \tilde{\beta}_j|^{-1}$ and $w_{jk(+)} = (1 + r_{jk})^{-1}|\tilde{\beta}_k + \tilde{\beta}_j|^{-1}$ for $1 \leq j < k \leq p$.

2.3 Geometric Interpretation of PACS Penalty

The geometric interpretation of the constrained least squares solutions illustrate the flexibility of the PACS penalty over the OSCAR penalty in accurately selecting the estimates. Aside from a constant, the contours of the least squares loss function are given by $(\beta - \hat{\beta}_{OLS})^T X^T X(\beta - \hat{\beta}_{OLS}) = K$. These contours are ellipses centered at the OLS solution. Since the predictors are standardized, when p=2 the principal axis of the contours are at $\pm 45^\circ$ to the horizontal. As the contours are in terms of $X^T X$, as opposed to $(X^T X)^{-1}$, positive correlation would yield contours that are at -45° whereas negative correlation give the reverse. In the (β_1, β_2) plane, intuitively, the solution in the first time the contours of the loss function hit the constraint region.

Figure 2 goes here.

Figure 2 illustrates the flexibility of the PACS approach over the OSCAR approach for a specific high correlation ($\rho = 0.85$) setup. In figures 2 (a) and 2 (b), we see that when the OLS solutions for β_1 and β_2 are close to each other, a specific case of the OSCAR penalty sets the OSCAR solutions as equal, $\hat{\beta}_1 = \hat{\beta}_2$, while the adaptive PACS penalty with weights given in Section 2.2.2 also sets $\hat{\beta}_1 = \hat{\beta}_2$. In figures 2 (c) and 2 (d) we see the same OSCAR and adaptive PACS penalty functions find different solutions when the OLS solution for β_1 is close to 0 and not close to that of β_2. Here the OSCAR solutions remain equal to each

other while the adaptive PACS sets $\hat{\beta}_1 = 0$. Thus the OSCAR solution is more dependent on the correlation of the predictors, and does not easily adapt to the different least squares solutions.

2.4 Computation and Implementation

2.4.1 Algorithm

The PACS estimates of (2) can be computed via quadratic programming with (p^2+p) parameters and $2p(p-1)+1$ linear constraints. As the number of predictors increases, quadratic programming becomes more computationally expensive and hence is not feasible for large, and even moderate p. In the following, we suggest an alternative computation technique that is more cost efficient. The algorithm is based on a local quadratic approximation of the penalty similar to that used in Fan and Li (2001). The penalty of the PACS in (2) can be locally approximated by a quadratic function. Suppose for a given value of λ, $P(|\beta_j|)$ is the penalty on the $|\beta_j|$ term of the penalty for $j = 1, ..., p$ and β_{0j} is a given initial value that is close to the minimizer of (2). If β_{0j} is not 0, using the first order expansion with respect to β_j^2, we have

$$P(|\beta_j|) = P((\beta_j^2)^{\frac{1}{2}}) \approx P(|\beta_{0j}|) + P'(|\beta_{0j}|)\ \frac{1}{2}(\beta_{0j}^2)^{-\frac{1}{2}}(\beta_j^2 - \beta_{0j}^2).$$

Thus $P(|\beta_j|) \approx P(|\beta_{0j}|) + \frac{1}{2}\frac{P'(|\beta_{0j}|)}{|\beta_{0j}|}(\beta_j^2 - \beta_{0j}^2)$ for $j = 1, ..., p$ and in the case of the PACS $|\beta_j| \approx \frac{\beta_j^2}{2|\beta_{0j}|} + K$, where K is a term not involving β_j and thus does not play a role in the optimization. Similarly, we approximate the penalty terms on the differences and sums of the coefficients. Thus both the loss function and penalty are quadratically expressed, and hence there is a closed form solution.

An iterative algorithm with closed form updating expressions follows from these local

quadratic approximations. Suppose, at step (t) of the algorithm, the current value, $\beta^{(t)}$, close to the minimizer of (2) is used to construct the following diagonal matrices; $I_w^{(t)} = diag(\frac{w_j}{|\beta_j^{(t)}|})$, $I_{w(-)}^{(t)} = diag(\frac{w_{jk(-)}}{|\beta_k^{(t)} - \beta_j^{(t)}|})$ and $I_{w(+)}^{(t)} = diag(\frac{w_{jk(+)}}{|\beta_j^{(t)} + \beta_k^{(t)}|})$. The value of the solution at step $(t+1)$ using these constructed matrices is given as

$$\hat{\beta}^{(t+1)} = (X^T X + \frac{\lambda}{2}(I_w^{(t)} + D_{(-)}^T I_{w(-)}^{(t)} D_{(-)} + D_{(+)}^T I_{w(+)}^{(t)} D_{(+)}))^- X^T \mathbf{y}.$$

The algorithm follows as:

1. Specify an initial starting value, $\beta^{(0)} = \hat{\beta}^{(0)}$ close to the minimizer of (2).

2. For step $(t+1)$, construct $I_w^{(t)}$, $I_{w(-)}^{(t)}$ and $I_{w(+)}^{(t)}$ and compute $\hat{\beta}^{(t+1)}$.

3. Let $t = t + 1$. Go to step 2 until convergence.

Proposition 2. *Assume $X^T X$ is full rank, then the optimization problem in (2) is strictly convex, and the proposed algorithm is guaranteed to converge to the unique minimizer. Note that if $X^T X$ is not full rank, then the problem is not strictly convex, and in that case it is guaranteed to converge to a local minimizer.*

The proof of Proposition 2 is due to the fact that it is an MM algorithm (see Hunter and Li, 2005).

Once the estimates are computed for a given λ, the next step is to select λ that corresponds to the best model. Cross-validation or an information criterion like AIC or BIC could be used for model selection. When using an information criterion, we need an estimate of the degrees of freedom. The degrees of freedom for the PACS is given as the number of unique absolute nonzero estimates as in the case of the OSCAR (Bondell and Reich, 2008). We use BIC to perform model selection in all examples in this paper.

3 Oracle Properties

The use of data adaptive weights in penalization assist in obtaining oracle properties for structure identification. Suppose $\tilde{\beta}$ is a $\sqrt{n}$-consistent estimator of β. Consider weights for this data adaptive PACS given by w^*, where w^* incorporates the weighting scheme from Section 2.2.2, so that w^* is given by $w_j^* = v_j|\tilde{\beta}_j|^{-1}$, $w_{jk(-)}^* = v_{jk(-)}|\tilde{\beta}_k - \tilde{\beta}_j|^{-1}$ and $w_{jk(+)}^* = v_{jk(+)}|\tilde{\beta}_k + \tilde{\beta}_j|^{-1}$ for $1 \leq j < k \leq p$. The choice of the constants v is flexible; one could use the correlation terms of Section 2.2.3 or any other choice satisfying the condition that $v_j \to c_j$, $v_{jk(-)} \to c_{jk(-)}$ and $v_{jk(+)} \to c_{jk(+)}$ with $0 < c_j, c_{jk(-)}, c_{jk(+)} < \infty$ for all $1 \leq j < k \leq p$. The adaptive PACS estimates are given as the minimizers of

$$\|\mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \beta_j\|^2 + \lambda_n \{\sum_{j=1}^{p} w_j^* |\beta_j| + \sum_{1 \leq j < k \leq p} w_{jk(-)}^* |\beta_k - \beta_j| \; + \sum_{1 \leq j < k \leq p} w_{jk(+)}^* |\beta_k + \beta_j|\}. \quad (3)$$

We now show that the adaptive PACS has the oracle property of variable selection and group identification. Consider the overparameterization, $\theta = M\beta$ from Section 2.2.1. Let $\mathcal{A} = \{i : \theta_i \neq 0, i = 1, ..., q\}$, $q = p^2$, denote the set of indices for which θ is truly non-zero. Let $\mathcal{A}_n$ denote the set of indices that are estimated to be non-zero.

Consider β^*, a vector of length $p_o \leq p$ that denotes the oracle parameter vector derived from $\mathcal{A}$. Let A^* be the $p_o \times p$ full rank matrix such that $\beta^* = A^*\beta$. For example, suppose the first two coefficients of β are truly equal in magnitude, then the first two elements of the first row of A^* would be equal to 0.5 and the remaining elements in the first row would equal 0. If a coefficient in β is not in the true model, then every element of the column of A^* corresponding to it would equal 0. We assume the regression model;

$$\mathbf{y} = X_o \beta^* + \epsilon,$$

where ϵ is a random vector with $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2 I$. Further assume that $\frac{1}{n} X_o^T X_o \to$

C, where X_o is the design matrix collapsed to the correct oracle structure determined by $\mathcal{A}$, and C is a positive definite matrix. The following theorem shows that the adaptive PACS has the oracle property.

Theorem 1. *(Oracle properties). Suppose* $\lambda_n \to \infty$ *and* $\frac{\lambda_n}{\sqrt{n}} \to 0$*, then the adaptive PACS estimates must satisfy the following:*

- *Consistency in structure identification:* $\lim_n P(\mathcal{A}_n = \mathcal{A}) = 1$.
- *Asymptotic normality:* $\sqrt{n}(A^*\hat{\beta} - A^*\beta) \to_d N(0, \sigma^2 C^{-1})$.

Remark 1. *We note here a correction to Theorem 1 in Bondell and Reich (2009). The proof of that theorem is also under the assumption that* $\lambda_n \to \infty$ *and* $\frac{\lambda_n}{\sqrt{n}} \to 0$.

4 Extension to Generalized Linear Models

We now study an extension of the PACS to GLMs and present a framework to show that an adaptive PACS is an oracle procedure. We consider estimation of penalized log likelihoods using an adaptive PACS penalty, where the likelihood belongs to the exponential family whose density is of the form $f(y|\mathbf{x}, \beta) = h(y)\exp(y(\mathbf{x}^T\beta - \phi(\mathbf{x}^T\beta)))$ (see McCullagh and Nelder, 1989). Suppose $\tilde{\beta}$ is the maximum likelihood estimate (MLE) in the GLM, then the adaptive PACS estimates are given as the minimizers of

$$\sum_{i=1}^{n}(-y_i(\mathbf{x}_i^T\beta) + \phi(\mathbf{x}_i^T\beta)) + \lambda_n\{\sum_{j=1}^{p} w_j^*|\beta_j| + \sum_{1\le j<k\le p} w_{jk(-)}^*|\beta_k - \beta_j| \ + \sum_{1\le j<k\le p} w_{jk(+)}^*|\beta_k + \beta_j|\},$$

where the weights are given as in Section 3. We assume the following regularity conditions:

(A) The Fisher information matrix for β^*, $I(\beta^*) = E[\phi''(\mathbf{x}_o^T\beta^*)\mathbf{x}_o\mathbf{x}_o^T]$, is positive definite.

(B) There is a sufficiently large open set $\mathcal{O}$ that contains β^* such that for all vectors of length p_o contained in $\mathcal{O}$, $|\phi'''(\mathbf{x}_o^T\beta^*)| \leq M(\mathbf{x}_o) < \infty$ and $E[M(\mathbf{x}_o)|\mathbf{x}_{oj}\mathbf{x}_{ok}\mathbf{x}_{ol}] < \infty$ for all $1 \leq j, k, l \leq p$.

Theorem 2. *(Oracle properties for GLMs). Assume conditions (A) and (B) and suppose* $\lambda_n \to \infty$ *and* $\frac{\lambda_n}{\sqrt{n}} \to 0$*, then the adaptive PACS estimates must satisfy the following:*

1. *Consistency in structure identification:* $\lim_n P(\mathcal{A}_n = \mathcal{A}) = 1$.

2. *Asymptotic normality:* $\sqrt{n}(A^*\hat{\beta} - A^*\beta) \to_d N(0, I(\beta^*)^{-1})$.

The proof of Theorem 2 follows the proof of Theorem 1 after a Taylor expansion of the objective function as done in the proof of Theorem 4 of Zou (2006).

4.1 Computation using Least Squares Approximation

The PACS estimates for GLMs can be computed via a Newton-Raphson type iterative algorithm. These algorithms are often computationally expensive and alternative methods if available are preferred. Recently, a novel method of computing an asymptotically equivalent solution was proposed by Wang and Leng (2007) where a least squares approximation (LSA) was applied to the negative log likelihood function, given by $-\frac{1}{n}\ell_n(\beta)$, and transforms the problem to its asymptotically equivalent Wald-statistic form. In particular, suppose the MLE, $\tilde{\beta}$ is $\sqrt{n}$-consistent, asymptotically normal with $\text{cov}(\tilde{\beta}) = \Gamma$, then the minimization of $-\frac{1}{n}\ell_n(\beta)$ is asymptotically equivalent to the minimization of $(\tilde{\beta} - \beta)^T\hat{\Gamma}^{-1}(\tilde{\beta} - \beta)$, where $\hat{\Gamma}$ is a consistent estimate of Γ. Consider the specific case that $\hat{\Gamma}^{-1}$ is symmetric and positive definite, then by the Cholesky decomposition we get $\hat{\Gamma}^{-1} = L^TL$, where L is an upper triangular matrix. So $(\tilde{\beta} - \beta)^T\hat{\Gamma}^{-1}(\tilde{\beta} - \beta) = (\tilde{\beta} - \beta)^TL^TL(\tilde{\beta} - \beta) = \|L(\tilde{\beta} - \beta)\|^2$. Then the minimization of $-\frac{1}{n}\ell_n(\beta)$ is asymptotically equivalent to the minimization of $\|L(\tilde{\beta} - \beta)\|^2$. Thus the asymptotic equivalent PACS estimates computed via the LSA are given as the

minimizers of

$$\|L(\tilde{\beta}-\beta)\|^2+\lambda\{\sum_{j=1}^{p}w_j|\beta_j|+\sum_{1\le j<k\le p}w_{jk(-)}|\beta_k-\beta_j|\ +\sum_{1\le j<k\le p}w_{jk(+)}|\beta_j+\beta_k|\}.$$

The PACS estimates can now be computed using the algorithm for the least squares model from Section 2.4. In this paper we suggest using the LSA technique to solve for the PACS estimates for GLMs.

5 Numerical Examples

5.1 Simulation Study

A simulation study compares the PACS approach with existing selection approaches in both prediction accuracy and model discovery. In all, four example are presented of 100 simulated data sets and different sample sizes given by n. The true models were simulated from a regression model given by $\mathbf{y}=X\beta+\epsilon,\ \epsilon\sim N(0,\sigma^2 I)$.

We compare two versions of the PACS approach with ridge regression, LASSO, adaptive LASSO and elastic net. The first PACS approach is the adaptive PACS (Adaptive PACS) with the weights given in Section 2.2.2 as $w_j=|\tilde{\beta}_j|^{-1}$, $w_{jk(-)}=|\tilde{\beta}_k-\tilde{\beta}_j|^{-1}$ and $w_{jk(+)}=|\tilde{\beta}_k+\tilde{\beta}_j|^{-1}$ for $1\le j<k\le p$. The second PACS approach is the adaptive PACS that incorporates correlations (AdCorr PACS) with the weights given in Section 2.2.3 as $w_j=|\tilde{\beta}_j|^{-1}$, $w_{jk(-)}=(1-r_{jk})^{-1}|\tilde{\beta}_k-\tilde{\beta}_j|^{-1}$ and $w_{jk(+)}=(1+r_{jk})^{-1}|\tilde{\beta}_k+\tilde{\beta}_j|^{-1}$ for $1\le j<k\le p$. Models were selected by BIC. We compare the methods in terms of model error (ME) and the resulting model complexity. Here the ME is given by $ME=(\hat{\beta}-\beta)^T V(\hat{\beta}-\beta)$, where V is the population covariance matrix of X. We report the median ME and its bootstrap standard error over 100 simulations. For model complexity, we report average degrees of freedom (DF), the percentage of correct models identified or selection accuracy (SA), the

percentage of correct groups identified or grouping accuracy (GA) and the percentage of both selection and grouping accuracy (SGA). Examples 3 and 4 have two clusters; here we additionally report grouping accuracy for each group given by GA-1 and GA-2 respectively. Note that none of the other methods perform grouping, so that their grouping accuracy will always be zero.

The first two examples are one cluster problems and the latter two examples have two clusters in the parameters. In Example 1, n=50, 100 and 200 observations are simulated with p=8 predictors. The true parameters are $\beta = (2, 2, 2, 0, 0, 0, 0, 0)^T$ and $\sigma = 1$. There are 3 important predictors and 5 unimportant predictors. The first three predictors form a cluster with pairwise correlation of 0.7. The remaining predictors are uncorrelated. All predictors are standard normal. Example 2 has the same setup as in Example 1, but here the covariance matrix of X is given by $\text{corr}(X_i, X_j) = 0.7^{|i-j|}$ for all $1 \leq i < j \leq p$. So in Example 2, the cluster of predictors is correlated with the unimportant predictors. We standardize all predictors before model fitting.

In Example 3, n=50, 100 and 200 observations are simulated with p=10 predictors. The true parameters are $\beta = (2, 2, 2, 1, 1, 0, 0, 0, 0, 0)^T$ and $\sigma = 1$. There are 5 important predictors and 5 unimportant predictors. The first three predictors form the first cluster with pairwise correlation of 0.7, the next two predictors form the second cluster with pairwise correlation of 0.7 while the remaining 5 predictors are uncorrelated. All predictors are standard normal. The setup for Example 4 is similar to that of Example 3, with the 5 important predictors given in Example 3, but the number of unimportant predictors in increased from 5 to 45. Again, the first three predictors form the first cluster with pairwise correlation of 0.7, the next two predictors form the second cluster with pairwise correlation of 0.7 while the remaining 45 predictors are uncorrelated. In Example 4 the sample sizes are set as n=200 and 400.

Table 1 summarizes the results for Examples 1 and 2. In Example 1, the two PACS

methods have the lowest ME for all sample sizes. In model complexity, the PACS methods have the lowest DF estimates. Although the elastic-net has the highest SA, we note that the elastic-net does not perform grouping. Only the two PACS methods successfully identify the cluster of predictors as seen in the GA and SGA columns. Here we also see that AdCorr PACS, which incorporates the pairwise correlation in the weighting scheme has lower ME and a higher rate of grouping than Adaptive PACS. We also observe that as the sample size increases the performance of the PACS methods improve. In Example 2, we notice very similar results as the results of Example 1. These results suggest that the PACS methods continues to identify the grouping structure successfully, even if the important predictors are correlated with the unimportant predictors.

Table 1 goes here.

Table 2 summarizes the results for Examples 3 and 4. In Example 3, the two PACS methods have the lowest ME for all sample sizes. In model complexity, the PACS methods have the lowest DF estimates. The PACS methods also have higher SA than the existing selection approaches, although for n=200, the adaptive lasso is slightly better in selection than AdCorr PACS. Here we also see that AdCorr PACS, which incorporates the pairwise correlation in the weighting scheme has lower ME and a higher rate of grouping than adaptive PACS. Example 4 is a tougher problem because of the increased number of unimportant predictors. For n=200, adaptive lasso and elastic-net have smaller ME than the Adaptive PACS method and also have better selection accuracy than both the PACS methods, although the AdCorr PACS method has the smallest ME. As the sample size increases to n=400, we notice that the adaptive lasso and the AdCorr PACS have the smallest ME. Although the PACS methods do not have the best selection accuracy among the selection methods, they exhibit improved grouping properties which are lacking in the competing selection methods. Also in example 4, we see that the assisted correlation weighting plays a significant role in grouping

accuracy with better results for AdCorr PACS than Adaptive PACS.

Table 2 goes here.

5.2 Analysis of Real Data Examples

In this section we illustrate the performance of the PACS approach with existing selection approaches in the analysis of examples of real data. Three data sets are studied, being the NCAA sports data from Mangold et al. (2003), the pollution data from McDonald and Schwing (1973) and the plasma data from Nierenberg et al. (1989). OLS, LASSO, adaptive LASSO, elastic net, adaptive PACS and AdCorr PACS as given in Section 5.1 were applied to the data. The predictors were standardized and the response was centered before performing analysis. Ridge regression estimates selected by AIC was used as adaptive weights for all data-adaptive approaches. Each data set is randomly split into a training and test set, where 20% of the data being used for testing purposes. The data set is randomly split 100 times each and models were selected on the training set using BIC. Test error (TE, average and s.e) of all methods are reported. The results are presented in table 3.

The NCAA sports data are taken from the 1996-99 editions of the US News "Best Colleges in America" and from the US Department of Education data which includes 97 NCAA Division 1A schools. The study aimed to show that successful sports programs raise graduation rates. The response is average 6 year graduation rate for 1996 through 1998 and the predictors are sociodemographic indicators and various sports indicators. The data contains 94 observations and 19 predictors and moderate pairwise correlations among the predictors, with 10% of the pairwise correlations being greater than 0.6 on the absolute scale. Table 3 shows that the PACS methods do significantly better than the existing selection approaches in test error. In fact, all the existing selection approaches perform worse than the OLS in prediction for this example.

The pollution data is from a study of the effects of various air pollution indicators and socio-demographic factors on mortality. The data contains 60 observations and 15 predictors and moderate pairwise correlations among the predictors, with 5% of the pairwise correlations being greater than 0.6 on the absolute scale. Table 3 shows that the PACS methods and the existing selection methods have comparable test error and all perform better than OLS in prediction. Of all the methods, Adaptive PACS is seen to have the lowest prediction error.

The plasma data taken is from a cross-sectional study to investigate the relationship between the predictors, personal characteristics and dietary factors, and the response, plasma concentrations of beta-carotene. The data contains 315 observations and 12 predictors and moderate pairwise correlations among the predictors, with 5% of the pairwise correlations being greater than 0.6 on the absolute scale. Table 3 shows that the PACS methods have the lowest prediction error among all the selection methods, while the existing selection methods have comparable or worse prediction error to OLS.

Table 3 goes here.

6 Discussion

In this paper we have proposed the PACS, a consistent group identification and variable selection procedure. The PACS produces a sparse model that clusters groups of predictive variables by setting their coefficients as equal, and in the process identifies these groups. Computationally, the PACS is shown to be easily implemented with an efficient computational strategy. Theoretical properties of the PACS are also studied and a data adaptive PACS is an oracle procedure for variable selection and group identification. The PACS is shown to have favorable predictive accuracy while identifying relevant groups in simulation studies and has favorable predictive accuracy in the analysis of real data.

References

Bondell, H. D. and Reich, B. J. (2008), "Simultaneous regression shrinkage, variable selection and clustering of predictors with OSCAR," *Biometrics*, 64, 115–123.

— (2009), "Simultaneous factor selection and collapsing of levels in ANOVA," *Biometrics*, 65, 169–177.

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

Fan, J. and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Hoerl, A. E. and Kennard, R. W. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55–67.

Hunter, D. R. and Li, R. (2005), "Variable Selection Using MM Algorithm," *Annals of Statistics*, 33, 1617–1642.

Mangold, W. D., Bean, L., and Adams, D. (2003), "The Impact of Intercollegiate Athletics on Graduation Rates Among Major NCAA Division I Universities," *Journal of Higher Education*, 70, 540–562.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models (2nd Ed.)*, New York: Chapman & Hall.

McDonald, G. C. and Schwing, R. C. (1973), "Instabilities of Regression Estimates Relating Air Pollution to Mortality," *Technometrics*, 15, 463–482.

Nierenberg, D. W., Stukel, T. A., Baron, J. A., Dain, B. J., Greenberg, E. R., and Group,

T. S. C. P. S. (1989), "Determinants of Plasma Levels of Beta-carotene and Retinol," *American Journal of Epidemiology*, 130, 511–521.

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

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society B*, 67, 91–108.

Tutz, G. and Ulbricht, J. (2009), "Penalized regression with correlation-based penalty," *Statistics and Computing*, 19, 239–253.

Wang, H. and Leng, C. (2007), "Unified LASSO Estimation by Least Squares Approximation," *Journal of the American Statistical Association*, 102, 1039–1048.

Wang, H., Li, G., and Jiang, G. (2007), "Robust Regression Shrinkage and Consistent Variable Selection Through the LAD-Lasso," *Journal of Business and Economic Statistics*, 25, 347–355.

Yuan, M. and Lin, Y. (2006), "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society B*, 68, 49–67.

Zhang, H. H. and Lu, W. (2007), "Adaptive Lasso for Coxs proportional hazards model," *Biometrika*, 94, 691–703.

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429.

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

Appendix

Proof of Proposition 1. Consider the matrix $M^- = \frac{1}{(2p-1)}[I_p \ D_{(-)}^T \ D_{(+)}^T]$. Then $M^-M = \frac{1}{(2p-1)}[I_p \ D_{(-)}^T D_{(-)} \ D_{(+)}^T D_{(+)}]$. Via direct calculation, one obtains $D_{(-)}^T D_{(-)} = pI_p - 1_p 1_p^T$ and $D_{(+)}^T D_{(+)} = (p-2)I_p + 1_p 1_p^T$, so $M^-M = I_p$. To show that M^- is the Moore-Penrose generalized inverse of M, it suffices to show

1. $M^-MM^- = M^-$.

2. $MM^-M = M$.

3. M^-M is symmetric.

4. MM^- is symmetric.

Clearly (1), (2) and (3) follow directly from the fact that $M^-M = I_p$. For (4),

$$MM^- = \begin{bmatrix} I_p & D_{(-)}^T & D_{(+)}^T \\ D_{(-)} & D_{(-)}D_{(-)}^T & D_{(-)}D_{(+)}^T \\ D_{(+)} & D_{(+)}D_{(-)}^T & D_{(+)}D_{(+)}^T \end{bmatrix}.$$

Clearly, MM^- is symmetric, and M^- is the Moore-Penrose generalized inverse of M. □

Proof of Theorem 1.1. We first prove for asymptotic normality. Let β_0 denotes the true parameter vector. Without loss of generality assume that $\beta_{j0} \geq 0$, for all $j = 1, 2, ..., p$. Then $\hat{u} = \arg\min V_n(u)$, where $\hat{\beta} = \beta_0 + \frac{\hat{u}}{\sqrt{n}}$ and

$$V_n(u) = u'(\frac{1}{n}X^TX)u - 2\frac{\epsilon' X}{\sqrt{n}}u + \frac{\lambda_n}{\sqrt{n}}P(u)$$

with

$$P(u) = \sum_{j}^{p} w_j^* \sqrt{n}(|\beta_{j0} + \frac{u_j}{\sqrt{n}}| - |\beta_{j0}|) + \sum_{1\le j<k\le p} w_{jk(-)}^* \sqrt{n}(|\beta_{k0} - \beta_{j0} - \frac{u_k - u_j}{\sqrt{n}}| - |\beta_{k0} - \beta_{j0}|)$$
$$+ \sum_{1\le j<k\le p} w_{jk(+)}^* \sqrt{n}(|\beta_{j0} + \beta_{k0} - \frac{u_j + u_k}{\sqrt{n}}| - |\beta_{j0} + \beta_{k0}|)$$

Consider the limiting behavior of $\frac{\lambda_n}{\sqrt{n}}P(u)$. By using arguments as in Zou, 2006 and Bondell and Reich, 2009, $\frac{\lambda_n}{\sqrt{n}}P(u)$ will go to zero for the correct structure and diverge under the incorrect structure. If $\beta_{j0} \neq 0$, $\beta_{k0} \neq 0$ and $\beta_{j0} \neq \beta_{k0}$, then $w_j^* \to_p c_j|\beta_{j0}|^{-1}$, $w_{jk(-)}^* \to_p c_{jk(-)}|\beta_{k0}-\beta_{j0}|^{-1}$ and $w_{jk(+)}^* \to_p c_{jk(+)}|\beta_{j0}+\beta_{k0}|^{-1}$. Also $\sqrt{n}(|\beta_{j0}+\frac{u_j}{\sqrt{n}}|-|\beta_{j0}|) \to u_j\ \mathrm{sgn}(\beta_{j0})$, $\sqrt{n}(|\beta_{k0}-\beta_{j0}-\frac{u_k-u_j}{\sqrt{n}}|-|\beta_{k0}-\beta_{j0}|) \to (u_k-u_j)\ \mathrm{sgn}(\beta_{k0}-\beta_{j0})$ and $\sqrt{n}(|\beta_{j0}+\beta_{k0}-\frac{u_j+u_k}{\sqrt{n}}|-|\beta_{j0}+\beta_{k0}|) \to (u_j+u_k)\ \mathrm{sgn}(\beta_{j0}+\beta_{k0})$. By Slutsky's theorem, we have $\frac{\lambda_n}{\sqrt{n}}P(u) \to_p 0$. If $\beta_{j0} = 0$ then $\sqrt{n}(|\beta_{j0}+\frac{u_j}{\sqrt{n}}|-|\beta_{j0}|) = |u_j|$ and $\frac{\lambda_n}{\sqrt{n}}w_j^* = \lambda_n v_j(|\sqrt{n}\hat{\beta}_j|)^{-1}$ where $\sqrt{n}\hat{\beta}_j = O_p(1)$, so $\frac{\lambda_n}{\sqrt{n}}P(u) \to \infty$ unless $\lambda_n\hat{u}_j \to 0$, since $\lambda_n \to \infty$. Similarly if $\beta_{j0} = \beta_{k0}$, $\sqrt{n}(|\beta_{k0} - \beta_{j0} - \frac{u_j-u_k}{\sqrt{n}}| - |\beta_{k0} - \beta_{j0}|) = |u_k - u_j|$ and $\frac{\lambda_n}{\sqrt{n}}w_{jk(-)}^* = \lambda_n v_{jk(-)}(|\sqrt{n}(\beta_{k0} - \beta_{j0})|)^{-1}$ where $\sqrt{n}(\hat{\beta}_k - \hat{\beta}_j) = O_p(1)$, so $\frac{\lambda_n}{\sqrt{n}}P(u) \to \infty$ unless $|u_k - u_j| \to 0$, since $\lambda_n \to \infty$. Furthermore, if $\mathcal{A}^c \subseteq \mathcal{A}_n^c$ then $V_n(u) = V_n^o(u_o)$, where $V_n^o(u_o)$ is the value of the objective function obtained after collapsing the design matrix X to the correct oracle structure determined by $\mathcal{A}$, i.e., $X = X_o$ and u_o is the corresponding oracle coefficient vector. Now by assumption, it follows that $\frac{1}{n}X_o'X_o \to C$ and $\frac{\epsilon' X_o}{\sqrt{n}} \to_d W \sim N(0, \sigma^2 C)$. Thus, we have $V_n(u) \to V(u)$ for every $\mathbf{u}$, where

$$\mathrm{V(u)} = \begin{cases} u_o' C u_o - 2u_o' W & \text{if } \mathcal{A}^c \subseteq \mathcal{A}_n^c\,, \\ \infty & \text{otherwise}\,. \end{cases}$$

Hence the asymptotic normality follows by noting that $V_n(u)$ is convex and the unique minimizer of V(u) is $C^{-1}W$. □

Proof of Theorem 1.1. We now show for consistency in structure identification. Note that

the asymptotic normality implies that $P(\mathcal{A}_n^c \cap \mathcal{A} \neq \emptyset) \to 0$. To complete the proof of consistency, we will show that $P(\mathcal{A}_n \cap \mathcal{A}^c \neq \emptyset) \to 0$. The proof is demonstrated through contradictions in 3 separate scenarios.

In the first scenario, let $\mathcal{A}^* = \{j : \beta_j \neq 0\}$ be the set of indices for coefficients that are truly non-zero and let $\mathcal{A}_n^*$ be the set of indices for coefficients that are estimated to be non-zero. Then if $\lambda_n \to \infty$ and $\tilde{\beta}$ is a $\sqrt{n}$-consistent estimator of β, $P(\mathcal{A}_n^* \cap \mathcal{A}^{*c} \neq \emptyset) \to 0$. We note that the asymptotic normality implies that $P(\mathcal{A}_n^{*c} \cap \mathcal{A}^* \neq \emptyset) \to 0$. Let $\mathcal{B}_n = \mathcal{A}_n^* \cap \mathcal{A}^{*c}$ be the set of indices corresponding to the coefficients that should be set to zero but were set to nonzero. We will show that $P(\mathcal{B}_n \neq \emptyset) \to 0$ i.e., that the true zeros will be set to zero with probability tending to one. Suppose $\mathcal{B}_n$ is not empty and also suppose $0 \leq \hat{\beta}_1 \leq ... \leq \hat{\beta}_p$. Also let $\hat{\beta}_m$ be the largest coefficient indexed by $\mathcal{B}_n$. Since $\hat{\beta}_m \neq 0$, (3) is differentiable w.r.t β_m and $\hat{\beta}_m$ satisfies

$$\frac{2}{\sqrt{n}} x_m^{'}(\mathbf{y} - X\hat{\beta}) = \frac{\lambda_n}{\sqrt{n}} \{w_m^* + \sum_{1 \leq j < m} w_{jm(-)}^* - \sum_{m < k \leq p} w_{mk(-)}^* + \sum_{1 \leq j < m} w_{jm(-)}^* + \sum_{m < k \leq p} w_{mk(-)}^*\}, \quad (4)$$

where x_m is the m^{th} column of X. Note that each term that diverges in the sums on the right hand side are nonnegative. We know that $\beta_m = 0$ due to the indexing and since $\tilde{\beta}$ is a $\sqrt{n}$-consistent estimator of β, we have $\sqrt{n}\tilde{\beta}_m = O_p(1)$. Therefore $\frac{w_m^*}{\sqrt{n}} = v_m(\sqrt{n}|\tilde{\beta}_m|)^{-1} = O_p(1)$. Consider $\frac{w_{jm(-)}^*}{\sqrt{n}} = v_{jm(-)}(\sqrt{n}|\tilde{\beta}_m - \tilde{\beta}_j|)^{-1}$ for $j < m$. If $\hat{\beta}_j$ is also indexed by $\mathcal{B}_n$, then $\beta_j = 0$, $\sqrt{n}(\tilde{\beta}_m - \tilde{\beta}_j) = O_p(1)$ and $\frac{w_{jm(-)}^*}{\sqrt{n}} = O_p(1)$. If $\hat{\beta}_j$ is not indexed by $\mathcal{B}_n$, then $|\tilde{\beta}_m - \tilde{\beta}_j| \to |\beta_j|$ and $\frac{w_{jm(-)}^*}{\sqrt{n}} \to 0$. Hence $\frac{w_{jm(-)}^*}{\sqrt{n}} = O_p(1)$ for $j < m$. Consider $\frac{w_{mk(-)}^*}{\sqrt{n}} = v_{mk(-)}(\sqrt{n}|\tilde{\beta}_k - \tilde{\beta}_m|)^{-1}$, since $\hat{\beta}_k$ is not indexed by $\mathcal{B}_n$, we have $|\tilde{\beta}_k - \tilde{\beta}_m| \to |\beta_k|$ and hence $\frac{w_{mk(-)}^*}{\sqrt{n}} \to 0$ for $m < k$. Consider $\frac{w_{mj(+)}^*}{\sqrt{n}} = v_{mj(+)}(\sqrt{n}|\tilde{\beta}_m + \tilde{\beta}_j|)^{-1}$ for $j \neq m$. If $\hat{\beta}_j$ is also indexed by $\mathcal{B}_n$, then $\beta_j = 0$, $\sqrt{n}(\tilde{\beta}_m + \tilde{\beta}_j) = O_p(1)$ and $\frac{w_{mj(+)}^*}{\sqrt{n}} = O_p(1)$. If $\hat{\beta}_j$ is not indexed by $\mathcal{B}_n$, then $|\tilde{\beta}_m + \tilde{\beta}_j| \to |\beta_j|$ and $\frac{w_{mj(+)}^*}{\sqrt{n}} \to 0$. Hence $\frac{w_{mj(+)}^*}{\sqrt{n}} = O_p(1)$ for all $j \neq m$. So the right hand side of (4) is $\lambda_n \times O_p(1) = O_p(\lambda_n)$, while the left hand side is $O_p(1)$ by the asymptotic normality. Now

since $\lambda_n \to \infty$, we have a contradiction and $P(\mathcal{B}_n \neq \emptyset) \to 0$.

For the second scenario, let $\mathcal{A}_{(-)} = \{(j,k) : \beta_j \neq \beta_k\}$ be the set of indices for differences in coefficients that are truly non-zero and let $\mathcal{A}_{n(-)}$ be the set of indices for these differences that are estimated to be non-zero. Then if $\lambda_n \to \infty$ and $\tilde{\beta}$ is a $\sqrt{n}$-consistent estimator of β, $P(\mathcal{A}_{n(-)} \cap \mathcal{A}^c_{(-)} \neq \emptyset) \to 0$. We note that the asymptotic normality implies that $P(\mathcal{A}^c_{n(-)} \cap \mathcal{A}_{(-)} \neq \emptyset) \to 0$. Let $\mathcal{B}_n = \mathcal{A}_{n(-)} \cap \mathcal{A}^c_{(-)}$ be the set of indices corresponding to the differences in coefficients that should be set to zero but were set to nonzero. We will show that $P(\mathcal{B}_n \neq \emptyset) \to 0$ i.e., that the true zeros will be set to zero with probability tending to one. Suppose $\mathcal{B}_n$ is not empty and also suppose $0 \leq \hat{\beta}_1 \leq ... \leq \hat{\beta}_p$. $\mathcal{B}_n = \mathcal{B}_{n1} \cup \mathcal{B}_{n2}$, where $\mathcal{B}_{n1}$ corresponds to the set $\{(\beta_j = \beta_k = 0) \cap (\hat{\beta}_j - \hat{\beta}_k \neq 0)\}$ and $\mathcal{B}_{n2}$ corresponds to the set $\{(\beta_j = \beta_k \neq 0) \cap (\hat{\beta}_j - \hat{\beta}_k \neq 0)\}$. $P(\mathcal{B}_{n1} \neq \emptyset) \to 0$ follows directly from the proof in the first scenario. We shall now prove $P(\mathcal{B}_{n2} \neq \emptyset) \to 0$.

Suppose $\hat{\beta}_m$ is the largest coefficient indexed by $\mathcal{B}_{n2}$ i.e. $m = \max\{k : (j,k) \in \mathcal{B}_{n2} \text{ for some } j\}$ and $\hat{\beta}_q$ is the smallest coefficient indexed by $\mathcal{B}_{n2}$ such that (q,m) $\in \mathcal{B}_{n2}$. Consider the reparameterization given by $X\beta = H\eta$, where $\eta_j = \beta_j - \beta_q$ for $j \neq q$ and $\eta_j = \beta_q$ for $j = q$. By assumption $\hat{\eta}_m \neq 0$ and $\eta_m = 0$, and that $\hat{\eta}_m - \hat{\eta}_j \geq 0$ for all (j,m) $\in \mathcal{B}_{n2}$. $|\beta_j| = \eta_j + \eta_q$ for $j \neq q$ and $|\beta_j| = \eta_q$ for $j = q$ for $j \leq k$ at the solution. Also for $j \leq k$ at the solution, we have $|\beta_k - \beta_j| = \eta_k - \eta_j$ for $(j \neq q, k \neq q)$, $|\beta_k - \beta_j| = \eta_k$ for $(j = q \leq k)$ and $|\beta_k - \beta_j| = -\eta_j$ for $(j \leq k = q)$. Also for $j \leq k$ at the solution, we have $|\beta_k + \beta_j| = \eta_k + \eta_j + 2\eta_q$ for $(j \neq q, k \neq q)$, $|\beta_k + \beta_j| = \eta_k + 2\eta_q$ for $(j = q \leq k)$ and $|\beta_k + \beta_j| = \eta_j + 2\eta_q$ for $(j \leq k = q)$. For this parametrization, for $j \leq k$ at the solution,

$$
\begin{aligned}
\hat{\eta} \;=\; & \arg\min_{\eta \in \mathcal{B}_{n2}} \|\mathbf{y} - H\eta\|^2 + \frac{\lambda_n}{\sqrt{n}} \Big\{ \sum_{j \neq q} w^*_j(\eta_j + \eta_q) + w^*_q \eta_q \\
& + \sum_{j \neq q, k \neq q, j<k} w^*_{jk(-)}(\eta_k - \eta_j) + \sum_{j=q \leq k} w^*_{qk(-)}\eta_k + \sum_{j \leq k=q} w^*_{jq(-)}\eta_k \\
& + \sum_{j \neq q, k \neq q, j<k} w^*_{jk(+)}(\eta_k + \eta_j + 2\eta_q) + \sum_{j=q \leq k} w^*_{qk(+)}(\eta_k + 2\eta_q) + \sum_{j \leq k=q} w^*_{jq(-)}(\eta_j + 2\eta_q) \Big\}
\end{aligned} \tag{5}
$$

Since $\hat{\eta}_m \neq 0$, (5) is differentiable w.r.t η_m at the solution. Consider this derivative in the neighborhood of the solution on which the coefficients that are set equal remain equal. That is, the differences remain zero. In this neighborhood, the terms involving (k,m) $\in \mathcal{A}^c_{n(-)} \cap \mathcal{A}^c_{(-)}$ can be omitted as these terms will vanish in the objective function. A contradiction will be obtained in this neighborhood of the solution. Due to the differentiability, the solution $\hat{\eta}$ must satisfy

$$\begin{aligned} \frac{2}{\sqrt{n}} h_m^{'}(\mathbf{y} - H\hat{\eta}) &= \frac{\lambda_n}{\sqrt{n}} \{ w^*_m + \sum_{k \neq m, (k,m) \in \mathcal{A}_{(-)}} (-1)^{[m \leq k]} w^*_{km(-)} + \sum_{k \neq m, (k,m) \in \mathcal{B}_{n2}} w^*_{km(-)} \quad (6) \\ &\quad + \sum_{k \neq m, (k,m) \in \mathcal{A}_{(-)}} w^*_{km(+)} + \sum_{k \neq m, (k,m) \in \mathcal{B}_{n2}} w^*_{km(+)} \}, \end{aligned}$$

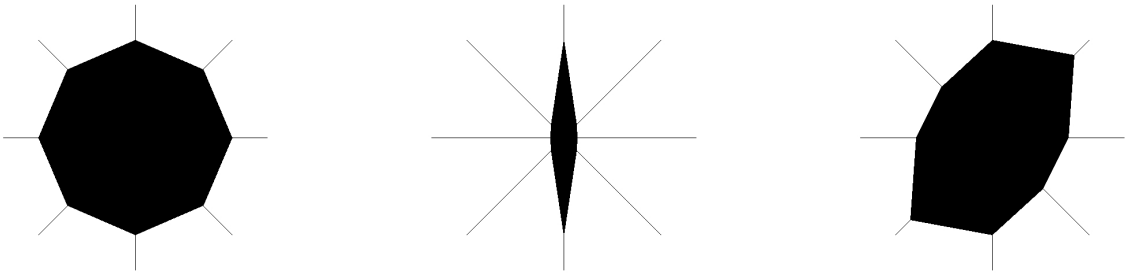
where h_m is the m^{th} column of H. Note that each term in the sums on the right hand side are nonnegative. Consider $\frac{w^*_m}{\sqrt{n}} = v_m(\sqrt{n}|\tilde{\beta}_m|)^{-1}$. If $\beta_m = 0$, we have $\frac{w^*_m}{\sqrt{n}} = O_p(1)$, and if $\beta_m \neq 0$, we have $|\tilde{\beta}_m| \to |\beta_m|$ and $\frac{w^*_m}{\sqrt{n}} \to 0$. Consider $(-1)^{[m \leq k]} \frac{w^*_{km(-)}}{\sqrt{n}} = (-1)^{[m \leq k]} v_{km(-)}(\sqrt{n}|\tilde{\beta}_m - \tilde{\beta}_k|)^{-1}$ for all $k \neq m, (k,m) \in \mathcal{A}_{(-)}$. Since (k,m)$\in \mathcal{A}_{(-)}$, $|\tilde{\beta}_m - \tilde{\beta}_k|^{-1} = O_p(1)$ and $(-1)^{[m \leq k]} \frac{w^*_{km(-)}}{\sqrt{n}} \to 0$ for all $k \neq m, (k,m) \in \mathcal{A}_{(-)}$. Consider $\frac{w^*_{km(-)}}{\sqrt{n}} = v_{km(-)}(\sqrt{n}|\tilde{\beta}_k - \tilde{\beta}_m|)^{-1}$ for all $k \neq m, (k,m) \in \mathcal{B}_{n2}$. Since (k,m)$\in \mathcal{B}_{n2}$, $(\sqrt{n}|\tilde{\beta}_k - \tilde{\beta}_m|)^{-1} = O_p(1)$ and $\frac{w^*_{km(-)}}{\sqrt{n}} = O_p(1)$ for all $k \neq m, (k,m) \in \mathcal{B}_{n2}$. Consider $\frac{w^*_{km(+)}}{\sqrt{n}} = v_{km(+)}(\sqrt{n}|\tilde{\beta}_k + \tilde{\beta}_m|)^{-1}$ for all $k \neq m, (k,m) \in \mathcal{A}_{(-)}$. Since (k,m)$\in \mathcal{A}_{(-)}$, $|\tilde{\beta}_k + \tilde{\beta}_m|^{-1} = O_p(1)$ and $\frac{w^*_{km(+)}}{\sqrt{n}} = O_p(\frac{1}{\sqrt{n}})$ for all $k \neq m, (k,m) \in \mathcal{A}_{(-)}$. Consider $\frac{w^*_{km(+)}}{\sqrt{n}} = v_{km(+)}(\sqrt{n}|\tilde{\beta}_k + \tilde{\beta}_m|)^{-1}$ for all $k \neq m, (k,m) \in \mathcal{B}_{n2}$. Since (k,m)$\in \mathcal{B}_{n2}$, $(\sqrt{n}|\tilde{\beta}_k + \tilde{\beta}_m|)^{-1} = O_p(1)$ and $\frac{w^*_{km(+)}}{\sqrt{n}} = O_p(1)$ for all $k \neq m, (k,m) \in \mathcal{B}_{n2}$. So the right hand side of (6) is $\lambda_n O_p(1) = O_p(\lambda_n)$, while the left hand side is $O_p(1)$ by the asymptotic normality. Now since $\lambda_n \to \infty$, we have a contradiction and $P(\mathcal{B}_{n2} \neq \emptyset) \to 0$.

For the third scenario, let $\mathcal{A}_{(+)} = \{(j,k) : \beta_j \neq -\beta_k\}$ be the set of indices for sums of coefficients that are truly non-zero and let $\mathcal{A}_{n(+)}$ be the set of indices for these sums that are estimated to be non-zero. Then if $\lambda_n \to \infty$ and $\tilde{\beta}$ is a $\sqrt{n}$-consistent estimator of β,

$P(\mathcal{A}_{n(+)} \cap \mathcal{A}^c_{(+)} \neq \emptyset) \to 0$. Let $\mathcal{B}_n = \mathcal{A}_{n(+)} \cap \mathcal{A}^c_{(+)}$ be the set of indices corresponding to the sums in coefficients that should be set to zero but were set to nonzero. We will show that $P(\mathcal{B}_n \neq \emptyset) \to 0$ i.e., that the true zeros will be set to zero with probability tending to one. The proof that $P(\mathcal{B}_n \neq \emptyset) \to 0$ follows directly from the proof in the first scenario. Suppose $\beta_j \geq 0$ and $\beta_k \geq 0$, then $\mathcal{B}_n$ is the set for which $\{\beta_j + \beta_k = 0, \hat{\beta}_j + \hat{\beta}_k \neq 0\}$. Since this set is contained in $\{\beta_j = 0, \hat{\beta}_j \neq 0\} \cup \{\beta_k = 0, \hat{\beta}_k \neq 0\}$, it suffices to show the property for the first scenario.

Together the three scenarios show that $P(\mathcal{A}_n \cap \mathcal{A}^c) \to 0$. As stated earlier, the proof of asymptotic normality implies that $P(\mathcal{A}^c_n \cap \mathcal{A}) \to 0$. Thus $\lim_n P(\mathcal{A}_n = \mathcal{A}) = 1$. □

Figure 1: Illustration to represent the flexibility of the PACS approach over the OSCAR approach in terms of the shaded constraint regions in the (β_1,β_2) plane.

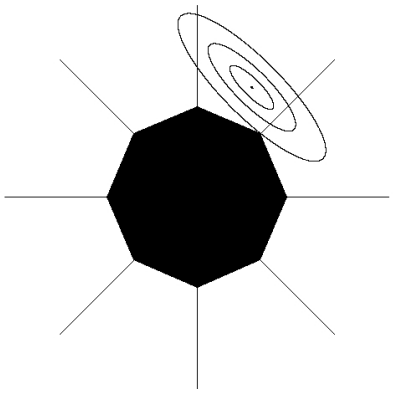


(a) OSCAR constraint region for a fixed value of c.

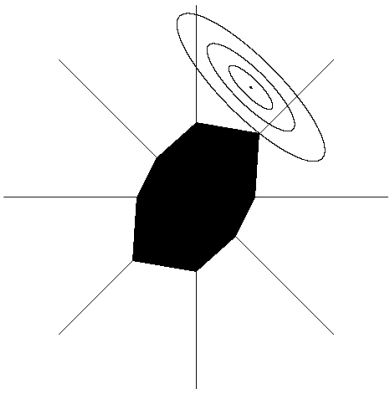
(b) PACS constraint region with one weighting scheme.

(c) PACS constraint region with a different weighting scheme.

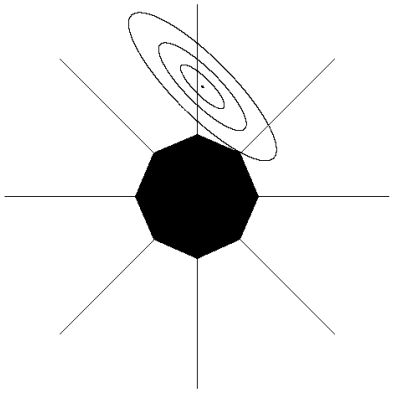
Figure 2: Graphical representation to represent the flexibility of the PACS approach over the OSCAR approach in the (β_1,β_2) plane. All figures represent correlation of $\rho = 0.85$. The top panel has OLS solution $\hat{\beta}_{OLS} = (1, 2)$ while the bottom panel has $\hat{\beta}_{OLS} = (0.1, 2)$. The solution is the first time the contours of the loss function hits the constraint region.



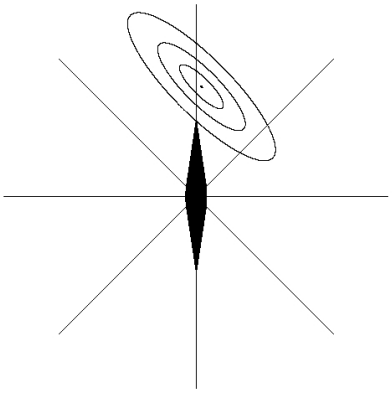
(a) When the OLS solutions of (β_1,β_2) are close to each other, OSCAR sets $\hat{\beta}_1 = \hat{\beta}_2$.



(b) When the OLS solutions of (β_1,β_2) are close to each other, PACS sets $\hat{\beta}_1 = \hat{\beta}_2$.



(c) When the OLS solutions of (β_1,β_2) are not close to each other and the OLS solution of β_1 is close to 0, OSCAR sets $\hat{\beta}_1 = \hat{\beta}_2$.



(d) When the OLS solutions of (β_1,β_2) are not close to each other and the OLS solution of β_1 is close to 0, PACS sets $\hat{\beta}_1 = 0$.

Table 1: Results for Example 1 and Example 2

Method	ME (s.e)	D.F	SA	GA	SGA
Example 1 (n=50)					
Ridge	0.1703(0.0082)	8.00	0	0	0
LASSO	0.1094(0.0093)	3.63	60	0	0
Adaptive LASSO	0.0840(0.0083)	3.36	74	0	0
Elastic Net	0.0695(0.0082)	3.32	78	0	0
Adaptive PACS	0.0617(0.0093)	2.04	67	44	32
AdCorr PACS	0.0496(0.0066)	1.64	66	77	56
Example 1 (n=100)					
Ridge	0.0762(0.0052)	8.00	0	0	0
LASSO	0.0454(0.0059)	3.43	68	0	0
Adaptive LASSO	0.0361(0.0045)	3.19	85	0	0
Elastic Net	0.0329(0.0048)	3.09	92	0	0
Adaptive PACS	0.0187(0.0034)	1.75	90	47	45
AdCorr PACS	0.0074(0.0017)	1.28	86	89	79
Example 1 (n=200)					
Ridge	0.0369(0.0029)	8.00	0	0	0
LASSO	0.0236(0.0018)	3.38	71	0	0
Adaptive LASSO	0.0120(0.0017)	3.10	93	0	0
Elastic Net	0.0113(0.0016)	3.04	97	0	0
Adaptive PACS	0.0056(0.0020)	1.51	91	63	59
AdCorr PACS	0.0022(0.0007)	1.24	88	91	82
Example 2 (n=50)					
Ridge	0.1618(0.0085)	8.00	0	0	0
LASSO	0.1178(0.0117)	3.85	51	0	0
Adaptive LASSO	0.0860(0.0099)	3.42	73	0	0
Elastic Net	0.0793(0.0084)	3.35	77	0	0
Adaptive PACS	0.0581(0.0097)	1.77	78	57	42
AdCorr PACS	0.0480(0.0075)	1.69	72	71	53
Example 2 (n=100)					
Ridge	0.0805(0.0059)	8.00	0	0	0
LASSO	0.0564(0.0061)	3.58	61	0	0
Adaptive LASSO	0.0334(0.0048)	3.13	88	0	0
Elastic Net	0.0372(0.0054)	3.11	89	0	0
Adaptive PACS	0.0176(0.0042)	1.57	92	58	56
AdCorr PACS	0.0090(0.0031)	1.35	88	80	76
Example 2 (n=200)					
Ridge	0.0368(0.0026)	8.00	0	0	0
LASSO	0.0232(0.0019)	3.41	69	0	0
Adaptive LASSO	0.0129(0.0016)	3.13	91	0	0
Elastic Net	0.0146(0.0018)	3.07	94	0	0
Adaptive PACS	0.0054(0.0016)	1.47	92	65	61
AdCorr PACS	0.0028(0.0007)	1.22	92	89	84

Table 2: Results for Example 3 and Example 4

Method	ME (s.e)	D.F	SA	GA-1	GA-2	GA	SGA
Example 3 (n=50)							
Ridge	0.1830(0.0158)	10.00	0	0	0	0	0
LASSO	0.1677(0.0128)	5.87	40	0	0	0	0
Adaptive LASSO	0.1316(0.0105)	5.32	70	0	0	0	0
Elastic Net	0.1330(0.0106)	5.42	65	0	0	0	0
Adaptive PACS	0.1210(0.0082)	3.60	72	41	45	20	16
AdCorr PACS	0.0873(0.0089)	2.98	73	72	64	48	38
Example 3 (n=100)							
Ridge	0.0960(0.0032)	10.00	0	0	0	0	0
LASSO	0.0810(0.0065)	5.69	57	0	0	0	0
Adaptive LASSO	0.0514(0.0041)	5.23	83	0	0	0	0
Elastic Net	0.0660(0.0062)	5.25	83	0	0	0	0
Adaptive PACS	0.0402(0.0040)	3.11	87	56	56	35	33
AdCorr PACS	0.0265(0.0040)	2.55	85	85	79	69	60
Example 3 (n=200)							
Ridge	0.0500(0.0023)	10.00	0	0	0	0	0
LASSO	0.0412(0.0036)	5.67	55	0	0	0	0
Adaptive LASSO	0.0240(0.0021)	5.09	93	0	0	0	0
Elastic Net	0.0289(0.0028)	5.15	89	0	0	0	0
Adaptive PACS	0.0148(0.0018)	2.83	95	72	55	37	35
AdCorr PACS	0.0111(0.0016)	2.23	92	95	90	85	78
Example 4 (n=200)							
Ridge	0.3067(0.0049)	50.00	0	0	0	0	0
LASSO	0.0723(0.0051)	6.09	38	0	0	0	0
Adaptive LASSO	0.0302(0.0018)	5.31	77	0	0	0	0
Elastic Net	0.0368(0.0022)	5.34	78	0	0	0	0
Adaptive PACS	0.0502(0.0085)	6.72	45	7	29	2	2
AdCorr PACS	0.0262(0.0029)	4.25	69	30	55	16	13
Example 4 (n=400)							
Ridge	0.1293(0.0041)	50.00	0	0	0	0	0
LASSO	0.0374(0.0019)	5.71	54	0	0	0	0
Adaptive LASSO	0.0121(0.0011)	5.15	89	0	0	0	0
Elastic Net	0.0178(0.0013)	5.12	91	0	0	0	0
Adaptive PACS	0.0166(0.0017)	4.56	72	24	34	8	8
AdCorr PACS	0.0121(0.0013)	3.38	81	51	60	33	31

Table 3: Results for Real Data Examples

Method	NCAA	Pollution	Plasma
	TE (s.e)	TE (s.e)	TE (s.e)
OLS	62.99 (1.62)	2052.25 (99.57)	0.52 (0.02)
LASSO	67.71 (1.45)	1480.16 (87.30)	0.52 (0.02)
Adaptive LASSO	64.51 (1.53)	1510.54 (81.82)	0.53 (0.02)
Elastic Net	67.63 (1.53)	1580.16 (84.78)	0.54 (0.02)
Adaptive PACS	57.49 (1.37)	1478.29 (79.99)	0.51 (0.02)
AdCorr PACS	57.34 (1.47)	1502.11 (78.12)	0.51 (0.02)