Generalized Correlation Shifting for Lasso

Izuru Miyazaki
The Graduate University for Advanced Studies

Hironori Fujisawa
The Institute of Statistical Mathematics
The Graduate University for Advanced Studies
RIKEN Center for Advanced Intelligence Project

Abstract

The Lasso has been widely used in a high-dimensional setting, but its estimation accuracy may become inadequate when the covariates are highly correlated or when the number of covariates is extremely large. To overcome this problem, we propose a novel preconditioner that adaptively induces a low-rank structure in the design matrix. The proposed preconditioner achieves a higher probability of sign correctness under some conditions. We establish theoretical guarantees showing that our method dominates the standard Lasso, and we further demonstrate its superiority over the correlation shifting. To validate its practical effectiveness, we conducted numerical experiments on synthetic and semi-real datasets, and the proposed method presented better performance than existing methods.

1 INTRODUCTION

In this paper, we consider the linear regression model in a high dimensional setting. The Lasso (Tibshirani, 1996) has become a canonical tool in such a setting, as its l_1 penalty simultaneously regularizes the model estimation and induces sparsity in the coefficients. However, the Lasso cannot present enough sign correctness when the covariates exhibit high correlations (Zhao and Yu, 2006), or when the number of covariates is much larger than the number of samples (Fan and Lv, 2008). To address this difficulty, there are two types of strategy.

(A) Refining the penalty. Several influential approaches focus on modifying the regularizer. The elastic net (Zou and Hastie, 2005) augments the l_1 penalty with the l_2 term. In addition to inducing sparsity in the estimation by the l_1 component, the l_2 component stabilizes estimation under multicollinearity and induces a grouping effect that tends to keep correlated covariates together, often improving both prediction and variable selection over Lasso when the covariates are highly correlated (Wang et al., 2020). The adaptive Lasso (Zou, 2006) employs data-dependent weights in the l_1 penalty, maintaining that the objective function is convex like Lasso, and achieves the oracle property (variable selection consistency and asymptotically normality with the oracle distribution). Concave penalties such as SCAD (Fan and Li, 2001) and MCP (Zhang, 2010) are designed to reduce estimation bias for large coefficients and also attain the oracle property.

(B) Preconditioning the data. The preconditioning constructs a new transformed data from the original data (X, Y) and then apply a penalized regression method to the transformed data. There are two representative approaches. One is the left-preconditioning, which uses a transformed data $(P_X X, P_y y)$ (Paul et al., 2008). For example, (Jia and Rohe, 2015) adopted a left-preconditioning to mitigate harmful correlations of covariates. The other is the right-preconditioning, which changes the basis in the parameter space so that the subsequent penalized regression is performed in a designed coordinate system (Kelner et al., 2022).

Apart from the two approaches discussed above, we also refer to the principal component regression (PCR) (Massy, 1965; Hotelling, 1957; Jolliffe, 1982). This method first constructs low-rank principal components as linear combinations of the covariates, then treats these components as new covariates, and finally makes a regression model using them. PCR is often employed to build good predictive models in a high dimensional setting (Bair et al., 2006; lldiko E. Frank and Friedman, 1993; Green and Romanov, 2025). However, when the objective is to recover the correct signs of

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

the original coefficients, the task becomes more challenging, since the original covariates are incorporated through these complex linear combinations.

In this paper, we propose a novel left-preconditioning technique with a low-rank structure, inspired by PCR. Our method selects an optimal low-rank left-preconditioner that maximizes the probability of sign correctness under some conditions. Our method can be regarded as a generalization of the correlation-shifting preconditioner (Huang and Jojic, 2011). To demonstrate the effectiveness of our proposed method, we present the following key contributions:

1. We theoretically show that the proposed method achieves a higher probability of sign correctness than the standard Lasso.
2. The ordinary correlation shifting loses its theoretical guarantee in some cases, in contrast the proposed method maintains the theoretical guarantee even in such cases.
3. We conducted numerical experiments on synthetic and semi-realistic datasets, demonstrating that the proposed method was better than existing approaches.

This paper is organized as follows. In Section 2, the Lasso and preconditioning are reviewed. In Section 3, to improve the sign correctness in the Lasso, we propose a novel left-preconditioning technique with a low-rank structure. In Section 4, some theoretical properties of the proposed method are presented, in particular, showing the superiority to the Lasso. In Section 5, the algorithm of the proposed method is provided. In Sections 6 and 7, numerical experiments on synthetic and semi-real datasets are illustrated. Section 8 presents the conclusion.

2 PRELIMINARIES

2.1 Lasso and its Sign Correctness

We begin with the following linear regression model:

$$y = X\beta^* + w, \tag{1}$$

where $y \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n\times p}$ is the design matrix, $\beta^* \in \mathbb{R}^p$ is the true coefficient vector, and $w \in \mathbb{R}^n$ is a Gaussian noise with $\mathbb{E}[w] = 0$ and $\mathrm{Var}[w] = \sigma^2 I_n$. We assume that β^* is s-sparse, i.e. $|\beta^*|_0 = s$, and denote its support by $S = \{j \in \{1,\dots,p\} \mid \beta^*_j \neq 0\}$. Let S^c be the complement of S.

The Lasso estimator of β^* is defined as

$$\hat{\beta} = \arg\min_{\beta\in\mathbb{R}^p} \left\{ \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \right\}.$$

In this paper, to evaluate the accuracy of variable selection, we focus on the sign correctness because it is amenable to theoretical analysis (Wainwright, 2009). Let the sign function be denoted by

$$\mathrm{sgn}(x) = \begin{cases} -1 & (x < 0) \\ 0 & (x = 0). \\ 1 & (x > 0) \end{cases}$$

The *sign correctness* of the estimator $\hat{\beta}$ is defined as

$$\mathrm{sgn}(\hat{\beta}) = \mathrm{sgn}(\beta^*).$$

This means that $\hat{\beta}$ has the same sign as the true parameter β^*.

We introduce the well-known lemma that characterizes the necessary and sufficient conditions for the sign correctness of the Lasso. For $i \in S$ and $j \in S^c$, define

$$Z_j = X_j^\top \left\{ X_S (X_S^\top X_S)^{-1} \mathrm{sgn}(\beta^*_S) + \Pi_{X_S^\perp}\left(\frac{w}{\lambda n}\right) \right\},$$
$$\Delta_i = e_i^\top \left(\frac{1}{n} X_S^\top X_S\right)^{-1} \left[\frac{1}{n} X_S^\top w - \lambda\, \mathrm{sgn}(\beta^*_S)\right],$$

where X_j is the j-th column of X, X_S is the submatrix of X consisting of columns indexed by S, β^*_S is the subvector of β^* consisting of entries indexed by S, $\Pi_{X_S^\perp}$ is the projection onto the orthogonal complement of the column space of X_S, and e_i is the i-th canonical base.

Lemma 1 (Wainwright, 2009)**.** *If $X_S^\top X_S$ is invertible, then the Lasso has a unique solution with sign correctness if and only if the following two events occur:*

$$\mathcal{M}(Z) = \left\{ \max_{j\in S^c} |Z_j| < 1 \right\}, \tag{2}$$

$$\mathcal{M}(\Delta) = \{\mathrm{sgn}(\beta^*_i + \Delta_i) = \mathrm{sgn}(\beta^*_i) \text{ for any } i \in S\}. \tag{3}$$

As in Lemma 1, we assume $X_S^\top X_S$ is invertible throughout this paper.

For later use, we introduce the following quantities:

$$\begin{aligned} \mu_j &= X_j^\top X_S (X_S^\top X_S)^{-1} \mathrm{sgn}(\beta^*_S), \\ \eta_j &= X_j^\top \Pi_{X_S^\perp}\left(\frac{w}{n}\right), \\ \epsilon_i &= e_i^\top \left(\frac{1}{n} X_S^\top X_S\right)^{-1} \frac{1}{n} X_S^\top w, \\ \gamma_i &= e_i^\top \left(\frac{1}{n} X_S^\top X_S\right)^{-1} \mathrm{sgn}(\beta^*_S). \end{aligned} \tag{4}$$

With these definitions, Z_j and Δ_i can be rewritten as

$$Z_j = \mu_j + \tfrac{\eta_j}{\lambda},\ \Delta_i = \epsilon_i - \lambda\gamma_i. \tag{5}$$

2.2 Preconditioning

In this paper, we focus on a left-preconditioner. The preconditioner aims at enhancing the performance of Lasso by transforming data (Paul et al., 2008). Given the original data (X, y), we apply preconditioners P_X and P_y to obtain transformed variables $\tilde{X} = P_X X$ and $\tilde{y} = P_y y$. The resulting dataset $(\tilde{X}, \tilde{y})$ is then used in place of the original one, leading to what is referred to as the *preconditioned Lasso*:

$$\tilde{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\tilde{y} - \tilde{X}\beta\|_2^2 + \tilde{\lambda}\|\beta\|_1 \right\}. \quad (6)$$

The preconditioner reshapes the geometry of the squared loss, and then the level sets and curvature of the squared loss are governed by $\tilde{X}^\top \tilde{X}$ instead of $X^\top X$. Jia and Rohe (2015) proposed the preconditoner called the Puffer Transformation. This is equivalent to decorrelating (whitening) the covariates in a low-dimensional setting, transforming the quadratic-loss contours from ellipses into circles, where $P_X = P_y$ is the whitening matrix of X. Huang and Jojic (2011) proposed the preconditioner called the correlation shifting. This preconditioning aims at reducing correlations among the covariates by removing an effect of large singular values of X, where $P_X = P_y$ is constructed from the left singular matrix of X corresponding to the smaller eigen values. Through experiments on synthetic data and semi-synthetic data, it was demonstrated that this preconditioning improved the accuracy of variable selection compared with methods such as the lasso, elastic net, and adaptive lasso Huang and Jojic (2011). Later, Wauthier et al. (2013) theoretically showed that, compared with the lasso, this preconditioner achieves sign consistency over a wider range of penalty parameters. The proposed method is an extension of the correlation shifting, as described later. It should be noted that, because the preconditioner also transforms noise, this effect may outweight the benefits of transforming the data.

3 PROPOSED METHODS

3.1 Main Idea

Let the SVD of X be denoted by $X = UDV^\top$, where $U \in \mathbb{R}^{n\times r}$ is the left singular matrix, $V \in \mathbb{R}^{p\times r}$ is the right singular matrix, and $D \in \mathbb{R}^{r\times r}$ is the diagonal matrix with strictly positive entries ordered in non-increasing order. Let K be a subset of $\{1, \ldots, r\}$. Define U_K and V_K as the submatrices containing the columns of U and V indexed by K, respectively, and define D_K as the principal submatrix obtained by selecting from D only those diagonal entries whose indices belong to K. We propose the preconditioner

$$F = U_K U_K^\top \quad \text{for } K \subseteq \{1, \ldots, r\},$$

and then we apply the Lasso to the transformed data $(\tilde{X}, \tilde{y}) = (FX, Fy)$.

Clearly, F is the orthogonal projection matrix onto the column space of U_K. This preconditioning is closely related to the idea of the correlation shifting (Huang and Jojic, 2011) because it is a special case of our framework with $K = \{k+1, \ldots, r\}$ for some positive integer k. In other words, the correlation shifting projects X onto the subspace spanned by the singular vectors corresponding to the smaller singular values.

However, there is no reason that the singular vector space corresponding to smaller sigular values are more important for the parameter estimation. From a methodological perspective, it is preferable that the preconditioner is adaptively determined according to the data. Motivated by this perspective, we refer to our method as the generalized correlation shifting (GCS), in contrast to Huang's original correlation shifting (OCS).

A fundamental issue in GCS is how to select the subset K. In the following, we propose a selecting method of K so as to maximize the probability of sign conrrectness of GCS under some conditions.

3.2 Some Conditions and Related Properties

In this subsection, we prepare some conditions and related properties. Let the SVD of X_S be denoted by $X_S = L_S M_S R_S^\top$, where $L_S \in \mathbb{R}^{n\times l}$ is the left singular matrix, $R_S \in \mathbb{R}^{s\times l}$ is the right singular matrix, and $M_S \in \mathbb{R}^{l\times l}$ is the diagonal matrix. For a matrix A, let $\mathrm{span}(A)$ be the space spanned by the column vectors of A. We assume the following condition on K.

Condition 1 (Active-set Preservation Condition)**.**

$$\mathrm{span}(L_S) \subseteq \mathrm{span}(U_K).$$

This condition implies that the column space of the true active set X_S remains unchanged after the preconditioner F is operated, more precisely, $FX_S = X_S$, and the regression structure also remains unchanged because $FX\beta^* = FX_S\beta_S^* = X_S\beta_S^*$. This condition was originally introduced in the study of OCS (Wauthier et al., 2013).

Let the preconditioned versions of the quantities intro-

duced in Equation (4) and Equation (5) be denoted by

$$\tilde{\mu}_j \coloneqq \tilde{X}_j^\top \tilde{X}_S (\tilde{X}_S^\top \tilde{X}_S)^{-1} \operatorname{sgn}(\beta_S^*),$$
$$\tilde{\eta}_j \coloneqq \tilde{X}_j^\top \Pi_{\tilde{X}_S^\perp} \left(\frac{\tilde{w}}{n}\right),$$
$$\tilde{\epsilon}_i \coloneqq e_i^\top \left(\frac{1}{n}\tilde{X}_S^\top \tilde{X}_S\right)^{-1} \frac{1}{n}\tilde{X}_S^\top \tilde{w},$$
$$\tilde{\gamma}_i \coloneqq e_i^\top \left(\frac{1}{n}\tilde{X}_S^\top \tilde{X}_S\right)^{-1} \operatorname{sgn}(\beta_S^*),$$
$$\tilde{Z}_j \coloneqq \tilde{\mu}_j + \frac{\tilde{\eta}_j}{\lambda},\ \tilde{\Delta}_i \coloneqq \tilde{\epsilon}_i - \lambda\tilde{\gamma}_i,$$

where $\tilde{w} = Fw$. We can see under Condition 1 that many of these remain unchanged and the only change in GCS is $\tilde{\eta}$ (Wauthier et al., 2013).

Proposition 1. *Under Condition 1, for any $i \in S$ and $j \in S^c$,*

$$\tilde{\mu}_j = \mu_j,\ \tilde{\epsilon}_i = \epsilon_i,\ \tilde{\gamma}_i = \gamma_i,\ \tilde{\Delta}_i = \Delta_i,$$
$$\tilde{\eta}_j = X_j^\top (I - L_S L_S^\top) U_K U_K^\top \frac{w}{n}.$$

The following irrepresentable condition (IC) is well known as a necessary and sufficient condition for the sign consistency of Lasso.

Condition 2 (Irrepresentable condition)**.**

$$\max_{j \in S^c} |\mu_j| < 1 - \eta,$$

where η is a positive constant.

Several sufficient conditions for IC are available, including small constant correlation, bounded correlation, and polynomial decay of correlations in the design matrix. See Zhao and Yu (2006) for details. IC requires that the correlations between the active covariates and inactive ones be sufficiently small. We assume Condition 1 and Condition 2 hold throughout the rest of this paper.

3.3 How to select K

We consider the maximization of sign correctness probability in GCS. The sign correctness holds when Equation (2) and Equation (3) occur for $\tilde{\Delta}_i$ and $\tilde{Z}_i$, as mentioned in Lemma 1. The probability of the former remains unchanged after the preconditioning since $\tilde{\Delta}_i = \Delta_i$ from Proposition 1. Hence, we focus on the probability of the latter, $\mathbb{P}(M(\tilde{Z}_j))$.

We have $|\tilde{\mu}_j| = |\mu_j| < 1 - \eta$ for all $j \in S^c$ from Condition 2, and then $|\tilde{Z}_j| < |\tilde{\mu}_j| + |\tilde{\eta}_j|/\lambda < 1 - \eta + |\tilde{\eta}_j|/\lambda$. Hence, if $|\tilde{\eta}_j| < \lambda\eta$, then $\mathcal{M}(\tilde{Z}_j)$ occurs. From Proposition 1, the variance of $\tilde{\eta}_j$ is

$$\begin{aligned}\operatorname{Var}[\tilde{\eta}_j] &= X_j^\top (I - L_S L_S^\top) U_K U_K^\top \frac{\sigma^2}{n^2} (I - L_S L_S^\top) X_j \\ &= \|U_K^\top (I - L_S L_S^\top) X_j\|_2^2 \frac{\sigma^2}{n^2}. \end{aligned} \tag{7}$$

By Gaussianity of the noise, $\tilde{\eta}_j$ follows a normal distribution with mean 0 and variance (7). Therefore, we can expect that as variance (7) is smaller, we have a higher probability of $\mathcal{M}(\tilde{Z})$. This motivates a selecting method of K in GCS by minimizing $\max_{j \in S^c} \operatorname{Var}[\tilde{\eta}_j]$ under Condition 1.

Definition 1 (GCS with max-criterion)**.**

$$K^{\max} \in \arg \min_{K \subseteq \{1,\dots,r\}} \max_{j \in S^c} \|U_K^\top (I - L_S L_S^\top) X_j\|_2^2$$
$$\text{s.t.} \quad \operatorname{span}(L_S) \subseteq \operatorname{span}(U_K).$$

Let $\tilde{\boldsymbol{\eta}} = (\tilde{\eta}_j)_{j \in S^c}$. Its covariance matrix is given by

$$\operatorname{Var}[\tilde{\boldsymbol{\eta}}] \coloneqq \frac{\sigma^2}{n^2} X_{S^c}^\top (U_K U_K^\top - L_S L_S^\top) X_{S^c}. \tag{8}$$

Thus, $K^{\max}$ can be regarded as the minimizer of the largest diagonal entry of $\operatorname{Var}[\tilde{\boldsymbol{\eta}}]$. Following the analogy with the optimal design theory (Kiefer and Wolfowitz, 1959), we also define D-, A-, and E-criteria for GCS, corresponding respectively to D-optimality (Kiefer and Wolfowitz, 1959), A-optimality (Kiefer, 1974), and E-optimality (Kiefer, 1974). In this paper, we basically use the max-criterion because it is useful. (For a detailed comparison of these criteria, see the Appendix.)

4 THEORETICAL ANALYSIS

4.1 Comparison with Lasso

We can show the following theorem, which implies that GCS (and OCS), for any K, exhibits a higher probablity of sign correctness than the standard Lasso.

Theorem 1. *Under Condition 1 and Condition 2,*

$$\mathbb{P}(\operatorname{sgn}(\tilde{\beta}) = \operatorname{sgn}(\beta^*)) \geq \mathbb{P}(\operatorname{sgn}(\hat{\beta}) = \operatorname{sgn}(\beta^*)).$$

Proof. The key step of the proof is to establish that $\operatorname{Var}[\boldsymbol{\eta}] - \operatorname{Var}[\tilde{\boldsymbol{\eta}}] \succeq O$ for any K. From this inequality and the Gaussianity of the noise, we have $\mathbb{P}(\mathcal{M}(\tilde{Z})) \geq \mathbb{P}(\mathcal{M}(Z))$. From this inequality, Proposition 1 and Lemma 1, we can show the theorem. For details, see the Appendix. □

4.2 Comparison with OCS

Let $\Phi : \mathbb{R}^{(p-k)\times(p-k)} \to \mathbb{R}$ be a criterion function, and let K_{opt} denote the minimizer of $\Phi(\mathrm{Var}[\tilde{\boldsymbol{\eta}}])$. It is important to note that, in general, K_{opt} cannot serve as a minimizer in the Loewner order, i.e. it does not hold that $\mathrm{Var}[\tilde{\boldsymbol{\eta}}_K] - \mathrm{Var}[\tilde{\boldsymbol{\eta}}_{K_{opt}}] \succeq O$ for all K (Kiefer, 1974). This makes it difficult to clearly say whether GCS is better than OCS. Nevertheless, under a certain condition, we can still establish a distinct advantage of GCS over OCS, which is formalized in the following theorem.

Theorem 2. *Without loss of generality, assume that the first s columns of X correspond to the active covariates, and the remaining $p-s$ columns correspond to inactive covariates. Suppose that Σ, the population covariance matrix of X, is block-diagonal, given by $\Sigma = \mathrm{diag}(\Sigma_S, \Sigma_{S^c})$, where Σ_S and Σ_{S^c} are the $s \times s$ and $(p-s)\times(p-s)$ matrices, respectively, and*

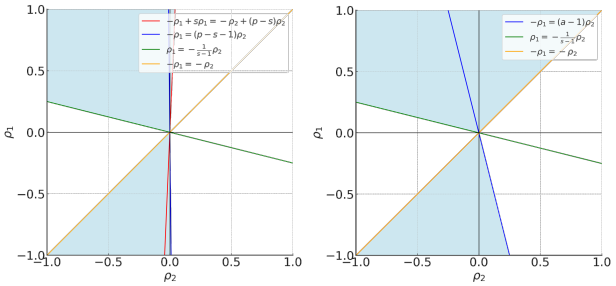
$$\lambda_{\max}(\Sigma_S) \geq \lambda_{\max}(\Sigma_{S^c}). \tag{9}$$

In addition, assume that the sample covariance matrix $\Sigma^{(n)} := X^\top X/n$ converges to Σ as $n \to \infty$. Then, in a limiting case, Condition 1 is satisfied for GCS with some K, but not satisfied for OCS.

Proof. Under the conditions of the theorem, the singular vectors eliminated by OCS lie in the column space of X_S, and therefore Condition 1 can no longer be satisfied. For details, see the Appendix. □

Example 1. *We illustrate two exmaples satisfying Equation (9). Let $\Lambda_a(\rho)$ denote the $a \times a$ matrix with ones on the diagonal and the constant ρ in the off-diagonal. The first example is $\Sigma_S = \Lambda_s(\rho_1)$ and $\Sigma_{S^c} = \Lambda_{p-s}(\rho_2)$. The parameter region where Equation (9) holds is shown in the left panel of Figure 1 for $p = 100$ and $s = 5$. The detail derivation is given in the Appendix. The second exmaple is a block-diagonal structure with $\Sigma_S = \Lambda_s(\rho_1)$ and $\Sigma_{S^c} = \mathrm{diag}(\Lambda_a(\rho_2), \ldots, \Lambda_a(\rho_2))$. For $s = 5$ and $a = 5$, the corresponding region is shown in the right panel of Figure 1. The second example yields a larger feasible region than the first one because Equation (9) reduces to $\lambda_{\max}(\Lambda_s(\rho_1)) \geq \lambda_{\max}(\Lambda_a(\rho_2))$ in the second exmaple, a weaker requirement than in the first example since $p - s \geq a$.*

Remark 1. *In the highlighted region of Example 1, OCS loses the theoretical guarantee provided in Theorem 1, in contrast GCS maintains the theoretical guarantee even in such regions.*



(a) $\Sigma_{S^c} = \Lambda_{p-s}(\rho_2)$ ($p = 100$, $s = 5$)

(b) $\Sigma_{S^c} = \mathrm{diag}(\Lambda_a(\rho_2), \ldots, \Lambda_a(\rho_2))$ ($s = 5$, $a = 5$)

Figure 1: Correlations satisfying the condition of Theorem 2. $\Sigma_S = \Lambda_s(\rho_1)$ in both cases.

5 IMPLEMENTATION

GCS cannot be directly performed because, in Definition 1, GCS is obtained through minimization involving an unknown active set S. Hence, by replacing S by an appropriate surrogate, we consider a surrogate problem instead of Definition 1.

First, we focus on the objective function in Definition 1. Let $\underline{S}$ be a set contained in S. This set can be obtained via a method such as the stability selection (Meinshausen and Bühlmann, 2010). By replacing S by $\underline{S}$ in the objective function, we consider the surrogate objective function

$$\max_{j\in\underline{S}^c} \|(U_K U_K^\top - L_{\underline{S}} L_{\underline{S}}^\top) X_j\|_2^2. \tag{10}$$

It can be verified that this serves as an upper bound of the original objective function, as shown below:

Proposition 2. *For any $\underline{S}(\subseteq S)$,*

$$\max_{j\in S^c} \|(U_K U_K^\top - L_S L_S^\top) X_j\|_2^2 \leq \max_{j\in\underline{S}^c} \|(U_K U_K^\top - L_{\underline{S}} L_{\underline{S}}^\top) X_j\|_2^2.$$

Proof. By a straightforward calculation, this proposition can be proved. For details, see the Appendix. □

Next, we consider the constraint in Definition 1. Let $\bar{S}$ be a set that contains S. This set can be obtained via a method such as the sure independence screening (Fan and Lv, 2008). By replacing S by $\bar{S}$ in the constraint, we have a larger set of K such that

$$\mathrm{span}(L_{\bar{S}}) \subseteq \mathrm{span}(U_K). \tag{11}$$

Finally, by integrating the surrogate objective function (10) with the surrogate constraint (11), we propose the following practical construction of $K^{\max}$:

Definition 2.

$$\hat{K}^{\max} \in \arg \min_{K \subseteq \{1,\dots,r\}} \max_{j \in \underline{S}^c} \|(U_K U_K^\top - L_{\underline{S}} L_{\underline{S}}^\top) X_j\|_2^2$$
$$\text{s.t.} \quad \text{span}(L_{\bar{S}}) \subseteq \text{span}(U_K).$$

Similarly, we can propose surrogate methods with other criteria. For issues related to feasible implementations of other criteria and their performance, please refer to the Appendix.

Since the minimization problem in Definition 2 is combinatorial, obtaining its exact solution is computationally expensive. In this paper, we employ a greedy method in which the size of K is gradually reduced by removing one element at a time. Let $\mathcal{L}(K; \underline{S})$ denote the objective function computed using the surrogate set $\underline{S}$. The greedy method is expressed in Algorithm 1.

Algorithm 1 Greedy GCS

Input: $X, \underline{S}, \bar{S}$
Output: GCS preconditioner F
$\tilde{K} \leftarrow \{1, \dots, \text{rank}(X)\}$; $L_{\text{old}} \leftarrow \mathcal{L}(\tilde{K}; \underline{S})$; $F \leftarrow I_n$
while $\text{span}(L_{\bar{S}}) \subseteq \text{span}(U_{\tilde{K}})$ **do**
  select $k \in \tilde{K}$ minimizing $\mathcal{L}(\tilde{K} \setminus \{k\}; \underline{S})$
    $L_{\text{new}} \leftarrow \mathcal{L}(\tilde{K} \setminus \{k\}; \underline{S})$
  if $L_{\text{new}} \leq L_{\text{old}}$ **then**
    $\tilde{K} \leftarrow \tilde{K} \setminus \{k\}$; $F = U_{\tilde{K}} U_{\tilde{K}}^T$
  else
    break
  end
end
return F

In Algorithm 1, an issue remains regarding how to properly verify whether the constraint $\text{span}(L_{\bar{S}}) \subseteq \text{span}(U_{\tilde{K}})$ is satisfied. To do this, using the R^2 value obtained by regressing each column of $L_{\bar{S}}$ onto the column space of $U_{\tilde{K}}$, we consider that the constraint $\text{span}(L_{\bar{S}}) \subseteq \text{span}(U_{\tilde{K}})$ is satisfied if R^2 is larger than a predefined threshold. This is not a rigorous criteion, but this flexibility is neccesary because $\bar{S}$, the surrogate estimator of S, has a variation generated from the data.

6 SIMULATION EXPERIMENT

6.1 Simulation Model and Evaluation Measure

The simulation model is

$$y = X\beta^* + w,$$

where y is the outcome, X is the design matrix, the i-th rows of X are i.i.d. samples drawn from $\mathcal{N}(0, I_p)$, and $w \sim \mathcal{N}(0, I_n)$. All nonzero entries of β^* are set to 1, and the size of the active set is fixed at 5. We set $n = 100$ and $p = 25, 200, 400, 600, 800, 1000$. Using this simulation model, we generated 30 random samples.

We obtained the GCS preconditioner F based on Algorithm 1 using (X, y), and set $\tilde{X} = FX$ and $\tilde{y} = Fy$. We performed the preconditioned Lasso using $(\tilde{X}, \tilde{y})$ based on Equation (6), and obtained the estimate $\hat{\beta}$. We determined the penalty parameter λ according to the OLS-BIC (Jia and Rohe, 2015).

Based on 30 random samples, each method was evaluated by Precison, Recall, and F1-measure, where Precison = TP/(TP + FP), Recall = TP/(TP + FN), F1-measure = 2/(1/Precision + 1/Recall) (TP: true positive, FP: false positive, FN: false negative). Each evaluation value was presented with the averages and standard deviations.

6.2 Known Active Set Setting

We examine the effectiveness of GCS in the ideal case where the true active set is known. We used the true active set only for constructing the preconditioner provided in Definition 1. In Figure 2, we compared GCS (R^2 thresholds of 0.3, 0.5, 0.7, 0.9) with Lasso and OCS ($k = 1, 5, 10$). GCS showed a similar performance to Lasso and OCS in low-dimensional settings ($n > p$); however, GCS demonstrated a better performance than Lasso and OCS in high-dimensional settings ($n < p$).

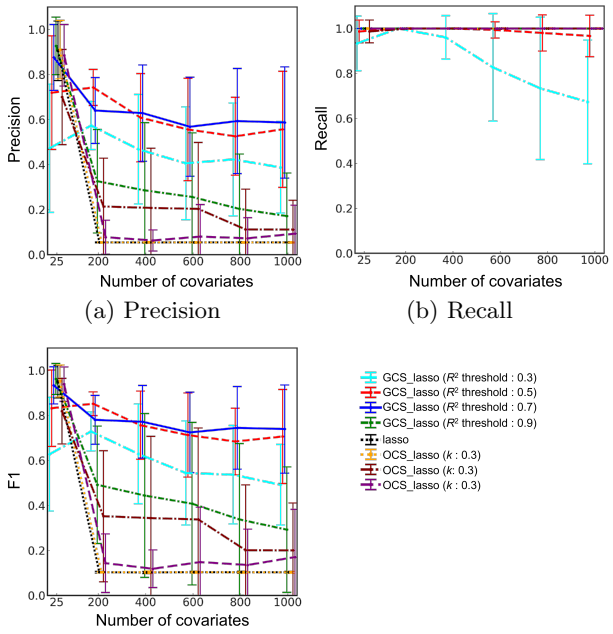
It should be noted that the R^2 threshold in GCS has a clear impact on the performance of GCS. The best performance was achieved around 0.7, and a larger value of the threshold does not lead to better performance. Later, we adaptively selected the threshold by OLS-BIC (Jia and Rohe, 2015).

Recall was almost 1 in most experiments. Hence, except in Figure 2, we report only the F1-measure because it is a monotone increasing function of Precision when Recall is fixed.

6.3 Influence of Using $\underline{S}$ and $\bar{S}$

First, we consider the case where S is replaced by a surrogate set $\underline{S}$ in the objective function and S is still known in the constraint ($\bar{S} = S$), and then we evaluate the influence of using $\underline{S}$, which was constructed by randomly selecting a specified proportion $r(= 0.2, 0.5, 0.8)$ of the indices from S. Except for this replacement, the experimental setting was the same as in the known active set setting. As shown in Figure 3, the influence of r was limited for any R^2 thresholds of GCS.

Next, we consider the case where S is replaces by a



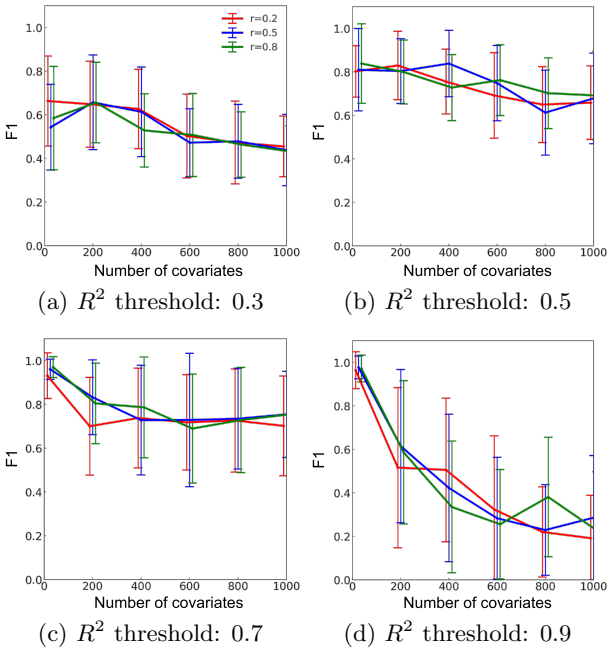
(a) Precision (b) Recall

(c) F1-measure

Figure 2: Precision, Recall, and F1 measure for simulation. Each evaluation value was presented with the averages and standard deviations.

surrogate set $\bar{S}$ in the constraint and S is still known in the objective function ($\underline{S} = S$), and then we evaluate the influence of using $\bar{S}$, which was constructed by augmenting S with a specified proportion $r(= 1, 3, 4)$ of randomly selected indices from S^c. Except for this replacement, the experimental setting was the same as in the known active set setting. As shown in Figure 4, the influence of r was limited for any R^2 thresholds of GCS.

Although it is difficult to theoretically evaluate how the accuracy of $\bar{S}$ and $\underline{S}$ affects the sign correctness of GCS, these results suggest that the performance of GCS is not highly sensitive to the accuracy of $\bar{S}$ and $\underline{S}$. To deepen our understanding of this, we examine how our choice of K is influenced by the estimation accuracy of S. We set the R^2 threshold to 0.7 and compare $\hat{K}^{\max}(S, S)$, which uses the true S for both the objective function and the constraint, with $\hat{K}^{\max}(\underline{S}, S)$, which uses $\underline{S}$ in the objective function, and $\hat{K}^{\max}(S, \bar{S})$, which uses $\bar{S}$ in the constraint. We calculated precision and recall by treating $\hat{K}^{\max}(S, S)$ as the ground truth for each value of r. The precisons and recalls of $\hat{K}^{\max}(\underline{S}, S)$ and $\hat{K}^{\max}(S, \bar{S})$ are shown in Table 1 and Table 2, respectively. Even when $r = 0.2$ for $\hat{K}^{\max}(\underline{S}, S)$, both precision and recall are high. Similarly, even when $r = 3$ for $\hat{K}^{\max}(S, \bar{S})$, both precision and recall are high. This indicates that the choice of K is weakly affected by the estimation accuracy of S. In addition, $\hat{K}^{\max}(\underline{S}, S)$ and $\hat{K}^{\max}(S, \bar{S})$ differed from those obtained by simply selecting smaller singular values, as in OCS, indicating that GCS constructs an efficient preconditioner by using the information contained in the estimated S.



(a) R^2 threshold: 0.3 (b) R^2 threshold: 0.5

(c) R^2 threshold: 0.7 (d) R^2 threshold: 0.9

Figure 3: Influence of using the surrogate set $\underline{S}$ with the different R^2 thresholds in GCS.

Table 1: Precision and recall of $\hat{K}^{\max}(\underline{S}, S)$ by treating $\hat{K}^{\max}(S, S)$ as the ground truth.

r	Precision	Recall
0.8	0.848 ± 0.037	0.807 ± 0.046
0.6	0.832 ± 0.039	0.806 ± 0.045
0.4	0.824 ± 0.053	0.800 ± 0.060
0.2	0.839 ± 0.032	0.821 ± 0.041

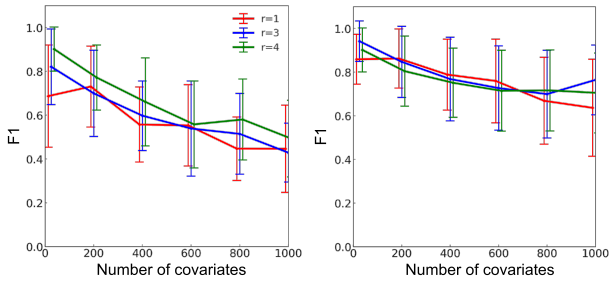
Table 2: Precision and recall of $\hat{K}^{\max}(S, \bar{S})$ by treating $\hat{K}^{\max}(S, S)$ as the ground truth.

r	Precision	Recall
0.5	0.865 ± 0.033	0.892 ± 0.040
1	0.873 ± 0.035	0.908 ± 0.031
2	0.856 ± 0.040	0.896 ± 0.033
3	0.845 ± 0.040	0.886 ± 0.036

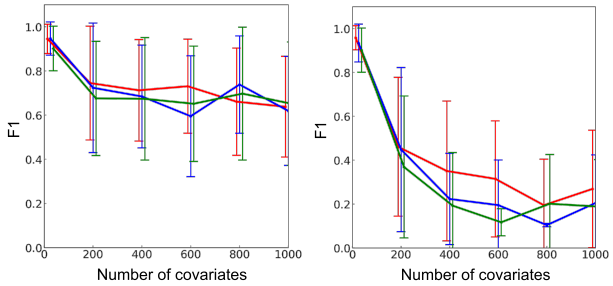
6.4 Unknown Active Set Setting

6.4.1 How to Make $\underline{S}$ and $\bar{S}$ in Practice

In Section 5, we suppose that $\underline{S} \subseteq S$ and $\bar{S} \supseteq S$. To construct $\underline{S}$ and $\bar{S}$, we employed the stability selection (SS) (Meinshausen and Bühlmann, 2010) and the sure independence screening (SIS) (Fan and Lv, 2008), respectively. SS provides control over the family-wise



(a) R^2 threshold: 0.3 (b) R^2 threshold: 0.5



(c) R^2 threshold: 0.7 (d) R^2 threshold: 0.9

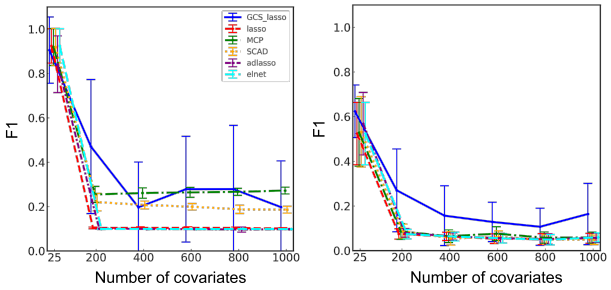
Figure 4: Influence of using the surrogate set $\bar{S}$ with the different R^2 thresholds in GCS.

error rate, and thus, with a sufficiently small selection threshold, the condition $\underline{S} \subseteq S$ is expected to hold. We used the SS implementation in `R` package `stabs` (Hofner and Hothorn, 2021) with the cutoff value 0.6. SIS enjoys the sure screening property, and thus $\bar{S} \supseteq S$ is expected to hold. We used the (non-iterative) SIS implementation in `R` package `SIS` (Saldana and Feng, 2018) with the default settings.

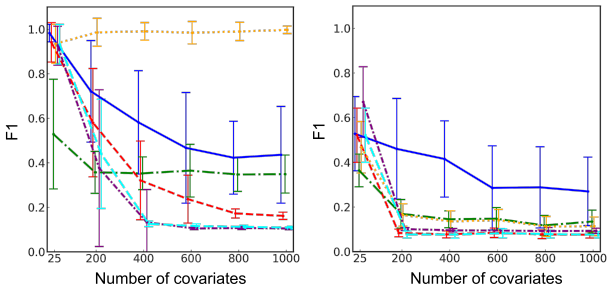
6.4.2 Comparison with Other Regularization Methods

We compared GCS with several representative regularization methods: Lasso, SCAD, MCP, elastic net (elnet), and adaptive lasso (adlasso). The tuning parameters of these methods were selected using OLS-BIC (Jia and Rohe, 2015). For the initial estimator in the adaptive lasso, we employed the ridge estimator with the tuning parameter determined by cross-validation. The puffer transformation (Jia and Rohe, 2015) was excluded from the comparison because the active set was estimated to be empty in most cases, especially in high-dimensional settings in preliminary studies.

The experimental settings were essentially the same as in the known active set setting, but we considered two types of design matrices: isotropic and block-independent. In the isotropic design, we assumed $X_i \sim$ i.i.d. $\mathcal{N}(0, I_p)$ for $i = 1, \ldots, n$. In the block-independent design, we assumed $X_i \sim$ i.i.d. $\mathcal{N}(0, \Sigma)$ for $i = 1, \ldots, n$, where $\Sigma = \mathrm{diag}(\Sigma_S, \Sigma_{S^c}), \Sigma_S = \Lambda_s(\rho), \Sigma_{S^c} = \Lambda_{p-s}(\rho)$. We set $\rho = 0.8$. All nonzero entries of β^* are set to 1 for a strong signal setting and 0.2 for a weak signal setting. The R^2 threshold of GCS was selected from $\{0.3, 0.4, \ldots, 0.9\}$ by OLS-BIC. The results are shown in Figure 5. GCS exhibited clearly higher performance when the signal was weak. When the signal was strong, GCS performed comparably to SCAD and MCP and was superior to the other methods in the isotropic design, and was inferior to SCAD but superior to the other methods in the block-independent design. We further compared our method with additional approaches. Refer to the Appendix for details.



(a) Isotropic design, $\beta^* = 1$ (b) Isotropic design, $\beta^* = 0.2$



(c) Block-independent design, $\beta^* = 1$ (d) Block-independent design, $\beta^* = 0.2$

Figure 5: F1-measure of various methods.

7 PERFORMANCE OF GCS FOR SEMI-REAL DATA

In this section, we examine the performance of GCS when evaluated using semi-real datasets. We treated six different real datasets of gene expression. A list of datasets used is given in Table 3. 'pomeroy' , 'west', 'subramanian', and 'alon' datasets were provided in `R` package `datamicroarray` (Ramhiser, 2012). In the same manner as in the known active set scenario, we generated 20 random samples of outcome. Many methods were applied to the semi-real datasets. The results are shown in Table 4. GCS was clearly better than the other methods except in some cases. GCS was slightly better in some cases (SCAD in the 'west' and

Table 3: Datasets.

Dataset	n	p
pomeroy (Pomeroy et al., 2002)	60	7128
west (West et al., 2001)	49	7129
subramanian (Subramanian et al., 2005)	50	10100
alon (Alon et al., 1999)	62	2000
golub (Golub et al., 1999)	72	3571
riboflavin (Bühlmann et al., 2014)	71	4088

'subramanian' datasets, Elastic Net and Lasso in the 'golub' dataset) and slightly inferior to SCAD in the 'riboflavin' dataset.

Table 4: F1-measures for semi-real data. Mean values are shown on the first line and the standard deviations are shown on the second line. The best F1-measure is indicated in bold for each setting.

GCS-lasso	MCP	SCAD	elnet	adlasso	lasso
pomeroy					
0.163	0.0659	0.0875	0.0934	0.0910	0.0962
(0.075)	(0.056)	(0.087)	(0.034)	(0.034)	(0.036)
west					
0.149	0.0767	0.133	0.104	0.0967	0.103
(0.109)	(0.0708)	(0.0968)	(0.0387)	(0.0314)	(0.0394)
subramanian					
0.154	0.119	0.131	0.0987	0.0946	0.101
(0.113)	(0.0814)	(0.114)	(0.0397)	(0.0419)	(0.0394)
alon					
0.152	0.1	0.11	0.09	0.0731	0.081
(0.056)	(0.079)	(0.0882)	(0.037)	(0.024)	(0.035)
golub					
0.201	0.048	0.029	0.199	0.120	0.199
(0.064)	(0.051)	(0.045)	(0.112)	(0.055)	(0.112)
riboflavin					
0.123	0.083	**0.135**	0.0903	0.073	0.0809
(0.093)	(0.071)	**(0.129)**	(0.037)	(0.024)	(0.0351)

8 CONCLUSION

In this paper, we have proposed a novel left-preconditioning technique with a low-rank structure, inspired by PCR. This is an extension of correlation shifting; hence it is called the generalized correlation shifting (GCS). GCS selects an optimal low-rank left-preconditioner that maximizes the probability of sign correctness under some conditions. Under certain assumptions, GCS theoretically dominates the standard Lasso, and further provides theoretical advantages over the original correlation shifting. Since some unknown information is necessary in GCS, we have constructed the surrogate optimization problem with the feasible algorithm from the dataset. We have conducted numerical experiments based on both synthetic and semi-real datasets. GCS demonstrated better performance than existing methods. **Limitation:** As our method depends on singular value decomposition, scalability can become a concern for large-scale dense matrices. To mitigate this, we note that preliminary feature screening and randomized SVD may be effective in reducing both runtime and memory requirements.

References

U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the USA*, 96(12):6745–6750, 1999.

E. Bair, T. Hastie, D. Paul, and R. Tibshirani. Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473):119–137, 2006.

P. Bühlmann, M. Kalisch, and L. Meier. High-dimensional statistics with a view towards applications in biology. *Annual Review of Statistics and Its Applications*, 1:255–278, 2014.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456): 1348–1360, 2001.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(5):849–911, 2008.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

A. Green and E. Romanov. The high-dimensional asymptotics of principal component regression, 2025.

B. Hofner and T. Hothorn. stabs: Stability selection with error control. `https://CRAN.R-project.org/package=stabs`, 2021. R package version 0.6-4.

H. Hotelling. The relations of the newer multivariate statistical methods to factor analysis. *British Journal of Statistical Psychology*, 10(2):69–79, 1957.

J. C. Huang and N. Jojic. Variable selection through correlation sifting. In *Research in Computational Molecular Biology*, pages 106–123. Springer, 2011.

J. Jia and K. Rohe. Preconditioning the lasso for sign consistency. *Electronic Journal of Statistics*, 9(1): 1150–1172, 2015.

I. T. Jolliffe. A note on the use of principal components in regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31(3):300–303, 1982.

J. A. Kelner, F. Koehler, R. Meka, and D. Rohatgi. On the power of preconditioning in sparse linear regression. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 550–561. IEEE, 2022.

J. Kiefer. General equivalence theory for optimum designs (approximate theory). *Annals of Statistics*, 2 (5):849–879, 1974.

J. Kiefer and J. Wolfowitz. Optimum designs in regression problems. *Annals of Mathematical Statistics*, 30 (2):271–294, 1959.

lldiko E. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

W. F. Massy. Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, 60(309):234–256, 1965.

N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(4):417–473, 2010.

D. Paul, E. Bair, T. Hastie, and R. Tibshirani. "preconditioning" for feature selection and regression in high-dimensional problems. *The Annals of Statistics*, 36(4):1595 – 1618, 2008.

S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. Kim, L. C. Goumnerova, P. M. Black, C. Lau, J. C. Allen, D. Zagzag, J. M. Olson, S. L. Curran, C.-S. Wetmore, J. A. Biegel, T. Poggio, E. S. Lander, J. P. Mesirov, T. R. Golub, and C. D. Stiles. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415(6870): 436–442, 2002.

J. Ramhiser. datamicroarray: Data sets for classification, clustering, and regression analysis with high-dimensional genomic data. `https://github.com/ramhiser/datamicroarray`, 2012. Accessed: 2025-10-02.

D. F. Saldana and Y. Feng. SIS: An R package for sure independence screening in ultrahigh-dimensional statistical models. *Journal of Statistical Software*, 83(2):1–25, 2018.

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the USA*, 102 (43):15545–15550, 2005.

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

M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l_1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202, 2009.

F. Wang, S. Mukherjee, S. Richardson, and S. M. Hill. High-dimensional regression in practice: an empirical study of finite-sample prediction, variable selection and ranking. *Statistics and Computing*, 30(3): 697–719, 2020.

F. L. Wauthier, N. Jojic, and M. I. Jordan. A comparative framework for preconditioned lasso algorithms. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

M. West, C. Blanchette, H. Dressman, E. Huang, S.-I. Lee, J. R. Spang, H. Zuzan, J. A. Olson, J. R. Marks, and J. R. Nevins. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences of the USA*, 98(20):11462–11467, 2001.

C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38: 894–942, 2010.

P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(90): 2541–2563, 2006.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101 (476):1418–1429, 2006.

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

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:

(a) Statements of the full set of assumptions of all theoretical results. [Yes]

(b) Complete proofs of all theoretical results. [Yes]

(c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:

(a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [No]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator If your work uses existing assets. [Yes]

(b) The license information of the assets, if applicable. [No]

(c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

(d) Information about consent from data providers/curators. [Not Applicable]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Generalized Correlation Shifting for Lasso

A1 Proof of Theorem 1

First, we present the following lemma concerning the probability that a random variable following a normal distribution is contained within a symmetric convex set:

Lemma A1 (Andereson's inequality)**.** *Assume $X_\Sigma \sim \mathcal{N}_d(0, \Sigma)$ and $\Sigma_1 \succeq \Sigma_2$, for a symmetric convex set $A \subset \mathbb{R}^d$,*

$$\mathbb{P}(X_{\Sigma_1} \in A) \leq \mathbb{P}(X_{\Sigma_2} \in A).$$

Proof. From Andereson's inequality (Anderson, 1996), for any $\boldsymbol{\mu} \in \mathbb{R}^d$,

$$\mathbb{P}(X \in A) \geq \mathbb{P}(X + \boldsymbol{\mu} \in A).$$

If the difference between two covariance matrices satisfies $\Sigma_1 - \Sigma_2 = CC^\top$, then

$$X_{\Sigma_1} \overset{\mathrm{d}}{=} X_{\Sigma_2} + CZ, \qquad Z \sim \mathcal{N}(0, I) \perp X_{\Sigma_2}.$$

For A,

$$\begin{aligned}
\mathbb{P}(X_{\Sigma_1} \in A) &= \mathbb{P}(X_{\Sigma_2} + CZ \in A) \\
&= \int_{\mathbb{R}^d} \mathbb{P}(X_{\Sigma_2} + Cz \in A | Z = z) f(z) dz \\
&= \int_{\mathbb{R}^d} \mathbb{P}(X_{\Sigma_2} + Cz \in A) f(z) dz \\
&\leq \mathbb{P}(X_{\Sigma_2} + Cz \in A) \\
&\leq \mathbb{P}(X_{\Sigma_2} \in A)
\end{aligned}$$

□

, where $f(z)$ is the probability density function of Z, and we used the Andereson's inequality in the last inequality.

Next, we define the counterpart of Equation (8) in the context of the Lasso as follows:

$$\boldsymbol{\eta} \coloneqq X_{S^c}^\top \Pi_{X_S^\perp} \left(\frac{w}{n}\right)$$

and

$$\mathrm{Var}[\boldsymbol{\eta}] \coloneqq \frac{\sigma^2}{n^2} X_{S^c}^\top (I - L_S L_S^\top) X_{S^c}.$$

From $I - U_K U_K^\top \succeq O$, $\mathrm{Var}[\boldsymbol{\eta}] \succeq \mathrm{Var}[\tilde{\boldsymbol{\eta}}]$. From IC, $|\boldsymbol{\mu}_j| < 1 - \eta$ and $\mathbb{P}(\mathcal{M}(Z)) = \mathbb{P}(|\tilde{\boldsymbol{\eta}}| < \lambda\eta)$. $A = \{\mathbf{a} \in \mathbb{R}^{p-s} : |a_i| \leq \lambda\eta\}$ is symmetric convex, then from Lemma A1,

$$\mathbb{P}(|\boldsymbol{\eta}| < \lambda\eta) \leq \mathbb{P}(|\tilde{\boldsymbol{\eta}}| < \lambda\eta),$$

and then

$$\mathbb{P}(\mathcal{M}(\tilde{Z})) \geq \mathbb{P}(\mathcal{M}(Z)).$$

In addition, from Proposition 1, we obtain $\mathbb{P}(\mathcal{M}(\tilde{\Delta})) = \mathbb{P}(\mathcal{M}(\Delta))$. Combining together, we finally obtain

$$\mathbb{P}(\mathrm{sgn}(\tilde{\beta}) = \mathrm{sgn}(\beta^*)) \geq \mathbb{P}(\mathrm{sgn}(\hat{\beta}) = \mathrm{sgn}(\beta^*)).$$

A2 Proof of Theorem 2

In OCS, the larger singular values of X are excluded. Consider the decomposition $X = UDV^\top$, where the diagonal entries of D are arranged in descending order. Since $\lambda_{\max}(\Sigma_s) \geq \lambda_{\max}(\Sigma_{p-s})$, the singular vector corresponding to the maximum singular value is the first column of U, denoted by U_1. If we further decompose X_S as $X_S = L_S M_S R_S^\top$, then by the block-dependency of X, it follows that $U_1 \in \mathrm{span}(L_S)$. Consequently, in OCS we have $U_1 \notin \mathrm{span}(U_K)$ for any $k \geq 1$. This implies that $\mathrm{span}(L_S) \not\subseteq \mathrm{span}(U_K)$, and hence Assumption 1 is no longer satisfied.

On the contrary, in the case of GCS, when $K = S$, we have $\tilde{\boldsymbol{\eta}} = \mathbf{0}$, and the minimum value of the objective function becomes 0 under any criterion. Therefore, a solution with $K = S$ exists for any criterion. In this case, Assumption 1 is naturally satisfied (with equality).

A3 Derivation of Figure 1 in Example 1

The eigenvalues of Σ_S are given by $1-\rho_1$ with multiplicity $s-1$, and $1-\rho_1+s\rho_1$ with multiplicity 1. Likewise, the eigenvalues of Σ_{S^c} are $1-\rho_2$ with multiplicity $p-s-1$, and $1-\rho_2+(p-s)\rho_2$ with multiplicity 1.

The conditions for $\lambda_{\max}(\Sigma_S) \geq \lambda_{\max}(\Sigma_{S^c})$:

1. in case $\rho_1 > 0$ and $\rho_2 > 0$, $(s-1)\rho_1 > (p-s-1)\rho_2$,
2. in case$\rho_1 > 0$ and $\rho_2 < 0$, $-\rho_1 > (p-s-1)\rho_2$,
3. in case $\rho_1 < 0$ and $\rho_2 > 0$, $(s-1)\rho_1 > -\rho_2$
4. in case $\rho_1 < 0$ and $\rho_2 < 0$, $-\rho_1 > -\rho_2$.

Thus, $\lambda_{\max}(\Sigma_s) \geq \lambda_{\max}(\Sigma_{p-s})$ holds in the highlighted region of (ρ_2, ρ_1) in Figure 1.

A4 Proof of Proposition 2

First we obtain the following lemma:

Lemma A2. *Let $\mathcal{U}, \mathcal{V}, \mathcal{W} \subset \mathbb{R}^d$ subspace such that $\mathcal{U} \subset \mathcal{V} \subset \mathcal{W}$, and denote each projection $P_\mathcal{U}, P_\mathcal{V}, P_\mathcal{W}$. Then*

$$\|(P_\mathcal{W} - P_\mathcal{V})x\|_2 \leq \|(P_\mathcal{W} - P_\mathcal{U})x\|_2 \qquad (\textit{for any } x \in \mathbb{R}^d).$$

Proof. Denote $w \coloneqq P_\mathcal{W} x$, $u \coloneqq P_\mathcal{U} x$, and $v \coloneqq P_\mathcal{V} x$. From $x - w \perp \mathcal{W} \supset \mathcal{V}$, $x - w \perp \mathcal{V}$, and then

$$v = P_\mathcal{V} x = P_\mathcal{V}(x - w + w) = P_\mathcal{V}(x - w) + P_\mathcal{V}(w) = P_\mathcal{V}(w)$$

Therefore, $w - v \perp \mathcal{V}$. Furthermore, from $v - u \in \mathcal{V}$,

$$\|w - u\|_2^2 = \|(w - v) + (v - u)\|_2^2 = \|w - v\|_2^2 + \|v - u\|_2^2 \geq \|w - v\|_2^2.$$

□

From the assumption, $\mathrm{span}(U_K) \supset \mathrm{span}(L_S) \supset \mathrm{span}(L_{\underline{S}})$ holds, and from Lemma A2, for any $x \in \mathbb{R}^n$,

$$\|(U_K U_K^\top - L_S L_S^\top)x\|_2^2 \leq \|(U_K U_K^\top - L_{\underline{S}} L_{\underline{S}}^\top)x\|_2^2. \tag{A1}$$

Finally, from $S^c \subset \underline{S}^c$,

$$\max_{j \in S^c} \|(U_K U_K^\top - L_{\underline{S}} L_{\underline{S}}^\top)X_j\|_2^2 \leq \max_{j \in \underline{S}^c} \|(U_K U_K^\top - L_{\underline{S}} L_{\underline{S}}^\top)X_j\|_2^2,$$

and we obtain

$$\max_{j \in S^c} \|(U_K U_K^\top - L_S L_S^\top)X_j\|_2^2 \leq \max_{j \in \underline{S}^c} \|(U_K U_K^\top - L_{\underline{S}} L_{\underline{S}}^\top)X_j\|_2^2.$$

A5 Other Criteria

A5.1 GCS with D-, A-, E-criterion

Definition A1 (GCS with D-, A-, E-criterion)**.** *Under the constraint of* $\mathrm{span}(L_S) \subseteq \mathrm{span}(U_K)$*,*

$$
\begin{aligned}
K^D &\in \arg\min_{K\subseteq\{1,\dots,r\}} \det[\mathrm{Var}[\tilde{\boldsymbol{\eta}}]],\\
K^A &\in \arg\min_{K\subseteq\{1,\dots,r\}} \mathrm{tr}[\mathrm{Var}[\tilde{\boldsymbol{\eta}}]],\\
K^E &\in \arg\min_{K\subseteq\{1,\dots,r\}} \|\mathrm{Var}[\tilde{\boldsymbol{\eta}}]\|_2.
\end{aligned}
$$

Upper bounds for the objective functions We introduce an upper bound of the objective function in the minimization problem for determining K in GCS max-criterion in Section 5. Similarly, we can propose a surrogate problem for K^A. We prepare a surrogate of S, $\underline{S}$ such that $\underline{S} \subseteq S$, and then consider the following surrogate minimization problem:

$$
\min_{K\subseteq\{1,\dots,r\}} \mathrm{tr}[X_{\underline{S}^c}^\top (U_K U_K^\top - L_{\underline{S}} L_{\underline{S}}^\top) X_{\underline{S}^c}].
$$

Again, it can be verified that this indeed serves as an upper bound of the original minimization problem, as shown below:

Proposition A1. *For any* $\underline{S}(\subseteq S)$*,*

$$
\begin{aligned}
&\mathrm{tr}[X_{S^c}^\top (U_K U_K^\top - U_S U_S^\top) X_{S^c}]\\
&\qquad \le \mathrm{tr}[X_{\underline{S}^c}^\top (U_K U_K^\top - U_{\underline{S}} U_{\underline{S}}^\top) X_{\underline{S}^c}]
\end{aligned}
$$

Proof. All diagonal entries of $\mathrm{Var}[\tilde{\boldsymbol{\eta}}]$ are non-negative. For any $j \in S^c \subseteq \underline{S}^c$, it is sufficient to show

$$
X_j^\top (U_K U_K^\top - L_S L_S^\top) X_j \le X_j^\top (U_K U_K^\top - L_{\underline{S}} L_{\underline{S}}^\top) X_j,
$$

and this is already shown in Equation (A1). □

Therefore, again combining with the alternative constraint that replace S with $\bar{S}$ such that $\bar{S} \supseteq S$ leads to the practical construction of K^A_{gcs} as follows:

Definition A2.

$$
\begin{aligned}
\hat{K^A_{gcs}} \in \arg\min_{K\subseteq\{1,\dots,r\}} &\ \mathrm{tr}[X_{\underline{S}^c}^\top (U_K U_K^\top - U_{\underline{S}} U_{\underline{S}}^\top) X_{\underline{S}^c}]\\
\text{s.t.} \quad &\ \mathrm{span}(U_{\bar{S}}) \subseteq \mathrm{span}(U_K).
\end{aligned}
$$

On the contrary, for the objective functions in K^D_{gcs} and K^E_{gcs}, with $\underline{S}$ such that $\underline{S} \subseteq S$,

$$
\det[X_{\underline{S}^c}^\top (U_K U_K^\top - U_{\underline{S}} U_{\underline{S}}^\top) X_{\underline{S}^c}]
$$

and

$$
\|X_{\underline{S}^c}^\top (U_K U_K^\top - U_{\underline{S}} U_{\underline{S}}^\top) X_{\underline{S}^c}\|_2
$$

are not the upper bounds in general of each objective function.

A5.2 Experiment Using Different Criteria

We compared the performance of gcs when using different criteria. The experimental setting was the same as in the known active set scenario in 6.1, and the R^2 threshold of GCS was fixed at 0.7. As shown in Figure A1, the overall performance remained broadly similar across different criteria. As mentioned earlier, for the E-criterion and D-criterion, it is not guaranteed that replacing S with $\underline{S}$ in each objective function will in general yield an upper bound of the original objective function. Nevertheless, at least in this experimental setting, the E-criterion and D-criterion appeared to achieve performance comparable to that of the other criteria.

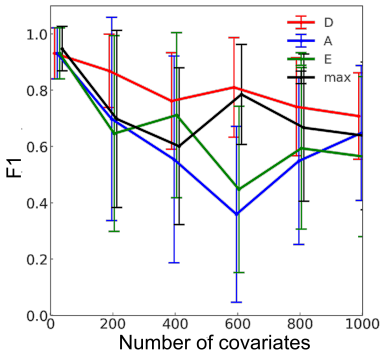
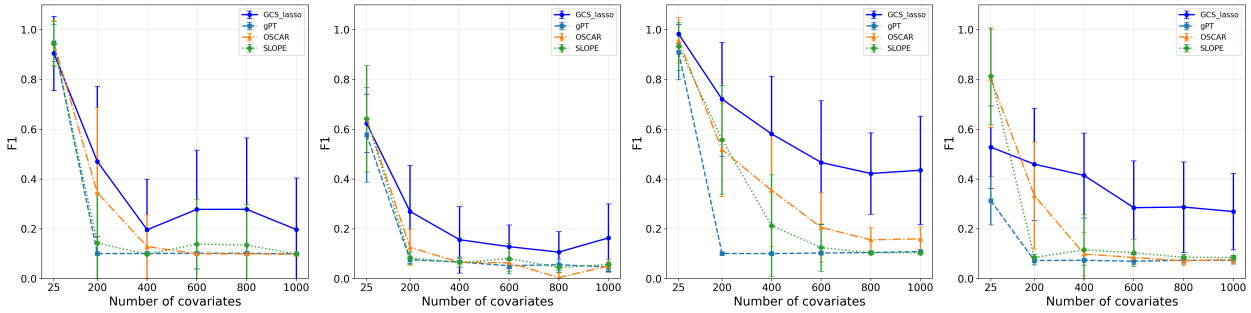


Figure A1: Influence of the criteria.

A6 Additional Experiments

We futher compared our method with additional approaches. These include generalized Puffer Transformation (gPT)(Jia and Rohe, 2015), OSCAR(Bondell and Reich, 2008), and SLOPE(Bogdan et al., 2015). We conducted experiments under the settings described in Section 6.4.2. The F1-measures for each method are presented in Figure A2. The proposed GCS outperformed these methods at most cases.



(a) Isotropic design, $\beta^* = 1$ (b) Isotropic design, $\beta^* = 0.2$ (c) Block-independent design, $\beta^* = 1$ (d) Block-independent design, $\beta^* = 0.2$

Figure A2: F1-measure of various methods.

References

T. W. Anderson. Some inequalities for symmetric sets with applications. *The Annals of Statistics*, 24(2):753–762, 1996.

M. Bogdan, E. van den Berg, C. Sabatti, W. Su, and E. J. Candès. SLOPE—Adaptive variable selection via convex optimization. *The Annals of Applied Statistics*, 9(3):1103 – 1140, 2015.

H. D. Bondell and B. J. Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, 64(1):115–123, 02 2008.

J. Jia and K. Rohe. Preconditioning the lasso for sign consistency. *Electronic Journal of Statistics*, 9(1):1150–1172, 2015.