Estimation of Linear Functionals in High Dimensional Linear Models: From Sparsity to Non-sparsity

Junlong Zhao, Yang Zhou, and Yufeng Liu*

Abstract

High dimensional linear models are commonly used in practice. In many applications, one is interested in linear transformations $\boldsymbol{\beta}^\top x$ of regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$, where x is a specific point and is not required to be identically distributed as the training data. One common approach is the plug-in technique which first estimates $\boldsymbol{\beta}$, then plugs the estimator in the linear transformation for prediction. Despite its popularity, estimation of $\boldsymbol{\beta}$ can be difficult for high dimensional problems. Commonly used assumptions in the literature include that the signal of coefficients $\boldsymbol{\beta}$ is sparse and predictors are weakly correlated. These assumptions, however, may not be easily verified, and can be violated in practice. When $\boldsymbol{\beta}$ is non-sparse or predictors are strongly correlated, estimation of $\boldsymbol{\beta}$ can be very difficult. In this paper, we propose a novel pointwise estimator for linear transformations of $\boldsymbol{\beta}$. This new estimator greatly relaxes the common assumptions for high dimensional problems, and is adaptive to the degree of sparsity of $\boldsymbol{\beta}$ and strength of correlations among the predictors. In particular, $\boldsymbol{\beta}$ can be sparse or non-sparse and predictors can be strongly or weakly correlated. The proposed method is simple for implementation. Numerical and theoretical results demonstrate the competitive advantages of the proposed method for a wide range of problems.

Keywords: Correlated predictors, eigenvalue sparsity, linear transformation, prediction

*Junlong Zhao is Professor, School of Statistics, Beijing Normal University, China, E-mail: zhaojunlong928@126.com. Yang Zhou is Assistant Professor, School of Statistics, Beijing Normal University, Beijing China, E-mail: bnu_zhouy@bnu.edu.cn. Yufeng Liu is Professor, Department of Statistics and Operations Research, Department of Genetics, Department of Biostatistics, Carolina Center for Genome Sciences, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, NC 27599, USA. E-mail: yfliu@email.unc.edu. The authors would like to thank the Editor, the Associate Editor, and reviewers, whose helpful comments and suggestions led to a much improved presentation. Junlong Zhao's research was supported in part by National Natural Science Foundation of China grants No. 11871104 and 12131006. Yang Zhou's research was supported in part by China Postdoctoral Science Foundation grants No. 2020M680226 and 2020TQ0014, and Youth Fund, No.310422112. Yufeng Liu's research was supported in part by US NIH grant R01GM126550 and NSF grants DMS2100729 and SES 2217440.

1 Introduction

With the advance of technology, high dimensional data are prevalent in many scientific disciplines such as biology, genetics and finance. Linear regression models are commonly used for the analysis of high dimensional data, typically with two important goals: prediction and interpretability. Variable selection can help to provide useful insights on the relationship between predictors and the response, and thus improve the interpretability of the resulting model. During the past several decades, many sparse penalized techniques have been proposed for simultaneous variable selection and prediction, including convex penalized methods (Tibshirani, 1996; Zou and Hastie, 2005), as well as nonconvex ones (Fan and Li, 2001; Zhang, 2010).

In this paper, we are interested in estimating linear transformations $\boldsymbol{\beta}^\top x$ of regression coefficients $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^\top \in \mathbb{R}^p$ for high dimensional linear models, where $x \in \mathbb{R}^p$ is a specific point and is not required to be from the same distribution as the training data. It relates to both coefficient estimation and prediction. For instance, sometimes we are interested in estimating β_1 and $\beta_1 - \beta_2$, where both of them can be expressed as $\boldsymbol{\beta}^\top x$ by taking x as $(1, 0, \ldots, 0)^\top$ and $(1, -1, 0, \ldots, 0)^\top$, respectively. On the other hand, for a typical prediction problem, x follows the same distribution as the training data.

To estimate $\boldsymbol{\beta}^\top x$, a natural and commonly used solution is to estimate $\boldsymbol{\beta}$ first by $\hat{\boldsymbol{\beta}}$ and construct the estimator $\hat{\boldsymbol{\beta}}^\top x$, which can be viewed as the *plug-in* one. The efficiency of the plug-in estimator depends on that of $\hat{\boldsymbol{\beta}}$. Despite its simplicity, obtaining a good estimate of $\boldsymbol{\beta}$ may not be easy in high-dimensional problems. If $\boldsymbol{\beta}$ is sparse (i.e. the support of $\boldsymbol{\beta}$, supp($\boldsymbol{\beta}$), is small), sparse regularized techniques such as the LASSO can be used to obtain a consistent estimator of $\boldsymbol{\beta}$. Desirable theoretical and numerical results have been established for various sparse penalized methods in the literature (see, for example, Bickel et al. (2009);

Raskutti et al. (2011); Bühlmann and Van De Geer (2011); Negahban et al. (2012)). These regularized methods assume that $\boldsymbol{\beta}$ is a sparse vector, which is difficult to verify in practice and may fail when supp($\boldsymbol{\beta}$) has a magnitude compatible with the sample size n or larger than n. The problem becomes more difficult when the predictors are strongly correlated since most sparse regularized methods work well on weakly dependent predictors.

We use a small simulation to illustrate the adverse effects of the sparsity degree of $\boldsymbol{\beta}$ on the plug-in estimator of $\boldsymbol{\beta}^\top x$ in the linear regression model (2.1), where X_i follows the normal distribution $N(0, \Sigma)$, $\boldsymbol{\beta} = \delta_0(\mathbf{1}_{p_0}^\top, \mathbf{0}_{p-p_0}^\top)^\top$ and $p_0 = r_0 p$, and $\Sigma = (\sigma_{ij})$, $\sigma_{ij} = 0.5^{|i-j|/\eta}$ with η controlling the correlation strength among the predictors. A larger value of r_0 implies a denser $\boldsymbol{\beta}$. The setup of δ_0 and other parameters are presented in Setting 1 of Section 5.1. The average testing errors of the plug-in estimators and our proposed PointWise Estimator (PWE) are shown in Figure 1. We can see that the errors of the plug-in estimators deteriorate quickly as r_0 increases. In contrast, our proposed estimator is much less sensitive to the change of the degree of non-sparsity.

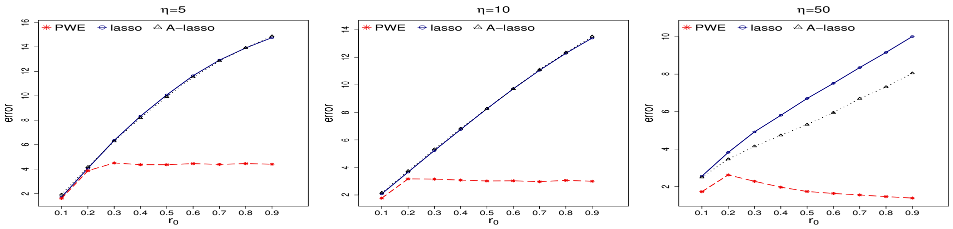

Figure 1: The effect of non-sparsity of $\boldsymbol{\beta}$ on plug-in estimators in terms of prediction error, where $p = 1000$. "A-lasso" and "lasso" denote the results of plug-in estimators with $\hat{\boldsymbol{\beta}}$ being adaptive LASSO and LASSO respectively, and PWE denotes the proposed method.

In typical prediction problems, a number of papers studied the convergence of prediction for various estimators, including LASSO, ridge, partial least squares, overparametrized estimators, and many others under different settings (Dalalyan et al., 2017; Zhang et al., 2017; Dobriban and Wager, 2018; Bartlett et al., 2020, etc.). It has been observed that LASSO and related methods are less affected by the correlation strength among predictors in prediction than in estimation problems (Hebiri and Lederer, 2013; Dalalyan et al., 2017). However, for some sparse vectors for x such as $x = (1, 0 \cdots, 0)^\top$, the estimation of $\boldsymbol{\beta}^\top x$ becomes that of the first coefficient and these methods are more affected by the correlation strength than prediction (Zou and Hastie, 2005). All the above mentioned methods consider the plug-in estimator and the average prediction error.

Different from these existing methods, we focus on $\boldsymbol{\beta}^\top x$ for a specific fixed x ($x \neq 0$) rather than on estimating $\boldsymbol{\beta}$ and the average prediction error, where x is not required to have the same distribution as the training data. The line of works that are closely related to ours are those on the hypothesis testing and confidence intervals of $\boldsymbol{\beta}^\top x$ in high dimensional linear models (van de Geer et al., 2014; Zhang and Zhang, 2014; Javanmard and Montanari, 2014; Lu et al., 2017; Cai and Guo, 2017; Zhu and Bradic, 2018, etc.). Most of these papers considered the case of $\boldsymbol{\beta}$ being ultra-sparse with $|\text{supp}(\boldsymbol{\beta})| \ll \sqrt{n}/\log p$. Cai and Guo (2017) considered the broader range where $|\text{supp}(\boldsymbol{\beta})|$ has an order no more than $n/\log p$. Zhu and Bradic (2018) considered the hypothesis testing and confidence intervals of $\boldsymbol{\beta}^\top x$ where $\boldsymbol{\beta}$ is allowed to be non-sparse by introducing a sparse auxiliary model, which can be restrictive. For example, if the predictor vector follows the normal distribution $N(0, \Sigma)$ and the sparse auxiliary model holds for any $x \in \mathbb{R}^p$ simultaneously, then Σ must be equal to I_p. Moreover, the estimator of $\boldsymbol{\beta}^\top x$ obtained from the confidence interval of Zhu and Bradic (2018), despite allowing $\boldsymbol{\beta}$ to be dense, works only when $p/n \to 0$ in prediction problems. Although these results have optimality in the minimax sense (Cai and Guo, 2017; Zhu and Bradic, 2018), they can be conservative and are actually determined by the most difficult case. In this paper, we introduce the sparsity of eigenvalues (or approximately low rank) of some matrices, which is shown to be complementary to the sparsity of $\boldsymbol{\beta}$. The most difficult case is actually the situation where both types of sparsity fail.

Our key observation is that we can directly target at $\gamma_x := \boldsymbol{\beta}^\top x$, treating it as an unknown parameter for estimation. We refer to the resulting estimate as the *pointwise* estimator. To this end, we propose a unified framework to leverage multiple sources of information. In many cases, the eigenvalues of the covariance matrix decrease dramatically, due to correlations among the predictors, which will be referred as *sparsity of eigenvalues* in the following descriptions. This type of sparsity is generally viewed as an adverse factor, making the estimation of $\boldsymbol{\beta}$ more difficult. Contrary to this popular view, we show that the sparsity of eigenvalues is beneficial in our framework and serves as a good complement to the sparsity of $\boldsymbol{\beta}$. In practice, two different kinds of test points x are of particular interest: (1) x is a given sparse vector, and (2) x is a random vector having the same distribution as the training data (i.e. the prediction problem). We give detailed results on these two special cases and compare our estimator with several other methods. The main contribution of this paper is that we propose a transformed model under a new basis, which provides a unified way to utilize different sources of information.

First, to utilize the sparsity of eigenvalues, we propose an estimator based on a basis consisting of eigenvectors of a specific matrix constructed from the training data. When the eigenvalues decrease at a certain rate, our estimator performs well for both kinds of test points x regardless of the sparsity of $\boldsymbol{\beta}$. On the other hand, if eigenvalues decrease slowly (or covariance matrix close to I_p), this estimator is less efficient; and consequently is inferior to LASSO when $\boldsymbol{\beta}$ is indeed sparse. In fact, the pointwise estimator using the sparsity of eigenvalues is complementary to LASSO.

Second, to leverage the information of $\boldsymbol{\beta}$, such as $\boldsymbol{\beta}$ being sparse, and the sparsity of eigenvalues jointly, we construct another basis based on an initial estimator $\hat{\boldsymbol{\beta}}$. It is shown that two types of information help each other: a faster decreasing rate of eigenvalues allows $\hat{\boldsymbol{\beta}}$ converging in a slower rate, and vice versa. When the test point x is a sparse vector, we show that the pointwise estimator performs well. The case of x being random as in prediction problems is more complicated in the sense that the sparsity degree of $\boldsymbol{\beta}$ should be taken into account. Hence, we consider a subset S_1 of $\{1, \cdots, p\}$ satisfying $S_1 \supseteq \text{supp}(\boldsymbol{\beta})$, where S_1 can be estimated from data. Specifically, we consider two cases: (1) Let $S_1 = \{1, \cdots, p\}$, which allows $\boldsymbol{\beta}$ to be sparse or dense. When sparsity of eigenvalues holds, our pointwise estimator performs well. When the eigenvalues are less sparse (or covariance matrix is close to I_p), our pointwise estimator performs similarly to the existing results on dense $\boldsymbol{\beta}$ in the literature. (2) When $\boldsymbol{\beta}$ is sparse, a smaller $|S_1|$ leads to a better estimator. If a good initial estimator $\hat{\boldsymbol{\beta}}$ and a good S_1 are available, our estimator's performance is similar to that of LASSO.

The rest of this paper is organized as follows. In Section 2, we propose our pointwise estimator for the linear transformation $\boldsymbol{\beta}^\top x$ in high dimensional linear models. Theoretical properties are established in Sections 3 and 4. Some simulated examples and real data analysis are presented in Section 5, followed by some discussions in Section 6. Proofs of the theoretical results are provided in the Supplementary Materials.

Notations. We first introduce some notations to be used for the paper. For any symmetric positive semidefinite matrix $A \in \mathbb{R}^{m \times m}$, denote the eigenvalues of A in a decreasing order as $\lambda_1(A) \geq \cdots \geq \lambda_m(A)$, and the smallest nonzero eigenvalues as $\lambda^+_{\min}(A)$. For any matrix $A \in \mathbb{R}^{m_1 \times m_2}$, $\lambda_{\max}(A), \lambda_{\min}(A)$ are the maximum and minimum singular values of A, respectively. For any vector $\boldsymbol{v} = (v_1, \cdots, v_m)^\top \in \mathbb{R}^m$, $\|\boldsymbol{v}\|$ and $\|\boldsymbol{v}\|_1$ denote the ℓ_2 and ℓ_1 norms of $\boldsymbol{v}$, respectively, and $\|\boldsymbol{v}\|_\infty = \max_{1 \leq j \leq m} |v_j|$; the support set of $\boldsymbol{v}$ is denoted as $\text{supp}(\boldsymbol{v})$. In addition, define $\|\boldsymbol{v}\|_A = \sqrt{\boldsymbol{v}^\top A \boldsymbol{v}}$ for any positive semidefinite matrix $A \in \mathbb{R}^{m \times m}$. For two sequences $\{a_n\}$ and $\{b_n\}$, both $a_n \lesssim b_n$ and $a_n = O(b_n)$ imply $\lim_n a_n/b_n \leq c$ for some constant $c < \infty$; both $a_n \gtrsim b_n$ and $a_n = \Omega(b_n)$ indicate that $\lim_n a_n/b_n \geq c$; $a_n \asymp b_n$ means that a_n has exactly the same order as b_n. For any integer i, let e_i denote the vector of zeros except the ith element being 1.

2 A unified framework for pointwise estimation

Suppose $(X_i, Y_i); 1 \leq i \leq n$, are *i.i.d.* from the following linear regression model

$$Y_i = X_i^\top \boldsymbol{\beta} + \epsilon_i; \quad 1 \leq i \leq n, \tag{2.1}$$

where $\epsilon_i \in \mathbb{R}$ satisfies $E(\epsilon_i) = 0$ and $\operatorname{var}(\epsilon_i) = \sigma^2 < \infty$, $X_i \in \mathbb{R}^p$ is independent of ϵ_i satisfying $E(X_i) = 0$, and $\operatorname{cov}(X_i) = \Sigma = (\sigma_{ij})$. Without loss of generality, we assume that $\sigma_{ii} = 1; i = 1, \ldots, p$, and that $\operatorname{var}(Y_i) < \infty$. Denote $\mathbf{X} = (X_1, \cdots, X_n)^\top \in \mathbb{R}^{n \times p}$, $\mathbf{Y} = (Y_1, \cdots, Y_n)^\top \in \mathbb{R}^n$, and $\varepsilon = (\epsilon_1, \cdots, \epsilon_n)^\top \in \mathbb{R}^n$. Then the model can be written as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$. Here the dimension p can be much larger than the sample size n. Let $x \in \mathbb{R}^p$ be a given point at which we intend to estimate $\boldsymbol{\beta}^\top x$. We assume that $X_i^\top x \neq 0$ for some $1 \leq i \leq n$, which can be checked numerically. Let $S_0 = \operatorname{supp}(\boldsymbol{\beta})$ of cardinality $s_0 = |S_0|$. Since a non-sparse $\boldsymbol{\beta}$ and the case where Σ might not be of full rank will be considered, we make the following identifiability condition and discuss some useful facts.

- When Σ is not of full rank, we assume that $\boldsymbol{\beta}$ falls into the column space of Σ for identifiability, due to the following reasons. Denote $X_i = \Sigma^{1/2}\tilde{X}_i; 1 \leq i \leq n$, where $\tilde{X}_i$ satisfies $E(\tilde{X}_i) = 0$ and $\operatorname{cov}(\tilde{X}_i) = I_p$. Let P_Σ be the projection matrix on the column space of Σ and $Q_\Sigma = I_p - P_\Sigma$. Then $X_i^\top \boldsymbol{\beta} = \tilde{X}_i^\top \Sigma^{1/2} \boldsymbol{\beta} = \tilde{X}_i^\top \Sigma^{1/2}(P_\Sigma + Q_\Sigma)\boldsymbol{\beta} = X_i^\top P_\Sigma \boldsymbol{\beta}$. Thus the parameter can be set as $P_\Sigma \boldsymbol{\beta}$, which falls into the column space of Σ.
- The magnitude of $\beta_j; j \in S_0$, depends on the sparsity degree s_0. Note that $\lambda_{\min}^+(\Sigma)\|\boldsymbol{\beta}\|^2 \leq \boldsymbol{\beta}^\top \Sigma \boldsymbol{\beta} < \operatorname{var}(Y_i) < \infty$, and consequently that $\|\boldsymbol{\beta}\|^2 \leq \operatorname{var}(Y_i)/\lambda_{\min}^+(\Sigma)$. Assume that β_j's with $j \in S_0$ are of the same magnitude. Then it follows that $|\beta_j| \lesssim [s_0 \lambda_{\min}^+(\Sigma)/\operatorname{var}(Y_i)]^{-1/2}$, $j \in S_0$. Particularly, if $\lambda_{\min}^+(\Sigma) \asymp 1$, $|\beta_j|$'s are of order $s_0^{-1/2}$, which can be small when s_0 is large.

Next we first introduce the transformed model based on a set of basis to leverage multiple sources of information in Section 2.1. The construction of basis is discussed in Section 2.2. A penalized estimator and a pointwise estimator are proposed in Section 2.3.

2.1 The transformed model

For any fixed $x \in \mathbb{R}^p$, let $P_x = xx^\top/\|x\|^2$ be the projection matrix on the space spanned by x and $Q_x = I_p - P_x$ be the projection matrix on the complementary space. Recall that $\gamma_x = \boldsymbol{\beta}^\top x$ and denote $\boldsymbol{\beta}_{Q_x} = Q_x\boldsymbol{\beta}$. Then one can write

$$\mathbf{X}\boldsymbol{\beta} = \mathbf{X}P_x\boldsymbol{\beta} + \mathbf{X}Q_x\boldsymbol{\beta} = \mathbf{X}x \cdot x^\top\boldsymbol{\beta}\|x\|^{-2} + \sqrt{n}\zeta_{\boldsymbol{\beta}} := \sqrt{n}z_x \cdot \alpha_x + \sqrt{n}\zeta_{\boldsymbol{\beta}}, \tag{2.2}$$

where $\alpha_x = \gamma_x \cdot \|x\|^{-2}\|\mathbf{X}x\|n^{-1/2} \in \mathbb{R}$, $z_x = \mathbf{X}x/\|\mathbf{X}x\| \in \mathbb{R}^n$ and $\zeta_{\boldsymbol{\beta}} = n^{-1/2}\mathbf{X}\boldsymbol{\beta}_{Q_x} \in \mathbb{R}^n$. Here α_x is a scaled version of γ_x such that the ℓ_2 norm of the predictor $\sqrt{n}z_x$ equals $\sqrt{n}$. Estimating γ_x is equivalent to that of α_x, since given α_x, one can compute γ_x directly from data $(\mathbf{X}, x)$. Then we get $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon = \sqrt{n}z_x\alpha_x + \sqrt{n}\zeta_{\boldsymbol{\beta}} + \varepsilon$, where $\zeta_{\boldsymbol{\beta}}$ is a nuisance parameter vector. Note that $\zeta_{\boldsymbol{\beta}}$ is a non-sparse vector in general, particularly when X_i's are *i.i.d.* variables; thus we have $n+1$ non-sparse parameters with the sample size n. To handle the difficulty, we introduce a set of basis $\Gamma \in \mathbb{R}^{n\times n}$ using different sources of information such that $\zeta_{\boldsymbol{\beta}}$ can be expressed sparsely under the set of basis.

The construction of Γ depends on the information at hand and will be elaborated in Section 2.2. For an invertible matrix $\Gamma \in \mathbb{R}^{n\times n}$, of which the columns are of unit length (i.e. $\Gamma_{\cdot j}^\top\Gamma_{\cdot j} = 1, 1 \le j \le n$), we denote $\sqrt{n}\zeta_{\boldsymbol{\beta}} = (\sqrt{n}\Gamma)(\Gamma^{-1}\zeta_{\boldsymbol{\beta}}) = \sqrt{n}\Gamma\boldsymbol{\theta}$, where $\boldsymbol{\theta} = \Gamma^{-1}\zeta_{\boldsymbol{\beta}} \in \mathbb{R}^n$. Although Γ here can be any invertible matrix, as shown later, the case we are interested in is Γ being (approximately) orthogonal. We hope that $\boldsymbol{\theta}$ is (approximately) sparse when Γ is chosen properly. Combining these together, we have the transformed linear model

$$\mathbf{Y} = \sqrt{n}z_x \cdot \alpha_x + \sqrt{n}\Gamma\boldsymbol{\theta} + \varepsilon = \mathbf{Z}\boldsymbol{\alpha} + \varepsilon, \tag{2.3}$$

where $\mathbf{Z} = \sqrt{n}(z_x, \Gamma) \in \mathbb{R}^{n\times(n+1)}$, $\boldsymbol{\alpha} = (\alpha_x, \boldsymbol{\theta}^\top)^\top \in \mathbb{R}^{n+1}$. The parameter $\boldsymbol{\theta}$ is treated as a

n-dimensional nuisance parameter vector. As shown later in Section 2.2, Γ plays a critical role in this model, providing a flexible way to leverage different sources of information. A naive choice is $\Gamma = I_n$ without using additional information, which will be discussed further in Section F of Supplementary Materials. In contrast to p parameters of the original linear model, the transformed model (2.3) has only $n+1$ unknown parameters that are (approximately) sparse. Without loss of generality, we assume that both $|\alpha_x|$ and consequently $\|\zeta_{\boldsymbol{\beta}}\|$ are bounded, given x and $\mathbf{X}$. This assumption is mild and holds in probability when x and X_i's are *i.i.d.* from $N(0, \Sigma)$. Detailed discussions are presented in Section A of the Supplementary Materials.

Denote by $\hat{\alpha}_x$ a generic estimator of α_x. Then γ_x can be estimated by $\hat{\gamma}_x = \hat{\alpha}_x \cdot \|x\|^2 \|\mathbf{X}x\|^{-1} n^{1/2}$. Given x, noting that $n^{-1}\|\mathbf{X}x\|^2 \to_p \|x\|_\Sigma^2$, where $\|x\|_\Sigma = \sqrt{x^\top \Sigma x}$, we see that the quantity $\|x\|^2/\|x\|_\Sigma$ affects the convergence rate of the estimator $\hat{\gamma}_x$. We investigate the magnitude of $\|x\|^2/\|x\|_\Sigma$, and consider two typical settings for clarity. Recall that x is a given vector, which may or may not have the same distribution as X_i.

Example 1. Let x be a sparse vector with the support set $S_x = \text{supp}(x)$ and the cardinality $|S_x| := s_x$. Examples of such x include $x = e_i$ or $x = e_i - e_j$. Denote $X_{iS_x} = (X_{ij}, j \in S_x)$. If the eigenvalues of $\Sigma_{S_x S_x} = \text{cov}(X_{iS_x})$ are both upper and lower bounded, and $\|x\|_\infty = O(1)$, then it follows that $\|x\|^2/\|x\|_\Sigma \asymp \|x\| \asymp s_x^{1/2}$.

Example 2. Let x be a random vector as in prediction problems. For simplicity, we assume x and X_i's are *i.i.d.* variables from $N(0, \Sigma)$. As shown in Section 4.2, it follows

$$\|x\|^2/\|x\|_\Sigma \asymp \sqrt{[tr(\Sigma)]^2/tr(\Sigma^2)} := M_\Sigma, \tag{2.4}$$

in probability. Moreover, it holds that $1 \le M_\Sigma \le p^{1/2}$ and M_Σ^2 can be viewed as the effective rank of Σ. Particularly, if some eigenvalues of Σ are much larger than the rest (e.g. the largest eigenvalue has the order p), then $M_\Sigma \asymp 1$; if Σ is close to I_p, then M_Σ is close to $p^{1/2}$ and γ_x is close to $p^{1/2}\alpha_x$.

Clearly, the magnitude of $\|x\|^2/\|x\|_\Sigma$ with a sparse x in Example 1 can be smaller than that of the dense x in Example 2. If x is not sparse, the sparsity of $\boldsymbol{\beta}$ can help as well. This observation motivates us to consider estimators utilizing the information of the sparsity degree of $\boldsymbol{\beta}$. For any subset $S_1 \subseteq \{1, \cdots, p\}$ such that $S_0 \subseteq S_1$, we observe that

$$\gamma_x = \boldsymbol{\beta}^\top x = \boldsymbol{\beta}_{S_0}^\top x_{S_0} = \boldsymbol{\beta}_{S_1}^\top x_{S_1} = \boldsymbol{\beta}^\top \tilde{x}_{S_1} = \gamma_{\tilde{x}_{S_1}},$$

where x_{S_0} and $\boldsymbol{\beta}_{S_0}$ are the subvectors of x and $\boldsymbol{\beta}$, respectively, and $\tilde{x}_{S_1}$ is a p-dimensional vector obtained by setting $x_{S_1^c} = 0$ in x. Thus, instead of estimating γ_x, one can equivalently consider prediction at the point $\tilde{x}_{S_1}$, which is a sparse vector when $|S_1|$ is small. Clearly, one can set $S_1 = \{1, \cdots, p\}$, and then $\gamma_{\tilde{x}_{S_1}}$ becomes γ_x. Estimating $\gamma_{\tilde{x}_{S_1}}$ provides a way to take the sparsity degree of $\boldsymbol{\beta}$ into account. In practice, one can choose different S_1's and select the best one by CV, as shown later in Section 2.3. In Section 4, we will compare our method with several existing methods in details for Examples 1 and 2.

Remark 1. *In Example 2 above, a smaller set S_1 is preferred. However, the true support set S_0 is unknown. When s_0 is small, choosing a set S_1 that covers S_0 is feasible. For example, S_1 can be taken as the support set of the LASSO estimator, or constructed by some screening methods (Fan and Lv, 2008).*

2.2 Construct basis utilizing multiple sources of information

Next we discuss the construction of Γ. For a positive semidefinite matrix $A \in \mathbb{R}^{p \times p}$, denote $\boldsymbol{\lambda}(A) = (\lambda_1(A), \cdots, \lambda_p(A))$ the vector of eigenvalues of A in a decreasing order. We say that $\boldsymbol{\lambda}(A)$ is approximately sparse when only a few eigenvalues are much larger than the average $p^{-1} \sum_{i=1}^{p} \lambda_i(A)$, and the detailed requirements on the decreasing rate of eigenvalues will be elaborated later. In practice, different sources of information may be available. For clarity of the presentation, we focus on two different sources of information.

Source I. We use the information of $\boldsymbol{\beta}$ through an initial estimator $\hat{\boldsymbol{\beta}}$. Note that a good estimator is available in some cases. In particular, if $\boldsymbol{\beta}$ is sparse, then $\hat{\boldsymbol{\beta}}$ can be obtained from that of LASSO, or other sparse regression methods. However, an estimator $\hat{\boldsymbol{\beta}}$ may not be good enough in many cases especially when $\boldsymbol{\beta}$ is less sparse and p is larger than n.

Source II. We utilize the dependence among predictors. Note that $\boldsymbol{\lambda}(n^{-1}\mathbf{X}Q_x\mathbf{X}^\top) = \boldsymbol{\lambda}(n^{-1}\mathbf{X}Q_xQ_x\mathbf{X}^\top)$ that equals the first n elements of $\boldsymbol{\lambda}(n^{-1}Q_x\mathbf{X}^\top\mathbf{X}Q_x)$, of which the population version is $\boldsymbol{\lambda}(Q_x\Sigma Q_x)$. When predictors are correlated such that $\boldsymbol{\lambda}(n^{-1}\mathbf{X}Q_x\mathbf{X}^\top)$ is (approximately) sparse, a key observation is that by choosing a suitable Γ, the parameter $\boldsymbol{\theta}$ can be (approximately) sparse regardless $\boldsymbol{\beta}$ being sparse or not (details are referred in Section 3.2). Thus it is feasible to estimate the parameters well in the transformed model (2.3).

Denote $\mathcal{C}_1 = \{\boldsymbol{\beta}\ \text{ is sparse}\}$, $\mathcal{C}_2 = \{\boldsymbol{\lambda}(n^{-1}\mathbf{X}Q_x\mathbf{X}^\top)\ \text{ is sparse}\}$, and let $\tilde{\mathcal{C}} = \mathcal{C}_1^c \cap \mathcal{C}_2^c$. The sparsity in $\mathcal{C}_2$ is complementary to that in $\mathcal{C}_1$. Both $\mathcal{C}_1$ and $\mathcal{C}_2$ are ideal cases with good estimators available (we will introduce estimators for Case $\mathcal{C}_2$ later), and the case $\tilde{\mathcal{C}}$ can be the least favorable case. There are intermediate cases between $\mathcal{C}_1$ and $\tilde{\mathcal{C}}$, when the degree of sparsity of $\boldsymbol{\beta}$ increases gradually. A similar argument applies between $\mathcal{C}_2$ and $\tilde{\mathcal{C}}$. To handle these complicated cases, it is natural to use these two different sources of information jointly. In fact, our approach works well under the complementary condition in Section 3.3 where stronger requirements on one type of sparsity weaken those of the other. It is possible that both $\mathcal{C}_1$ and $\mathcal{C}_2$ hold simultaneously, where our estimators leverage both sources of information.

Denote the spectral decomposition of $n^{-1}\mathbf{X}Q_x\mathbf{X}^\top$ as $\Gamma_{\text{eg}}\Psi\Gamma_{\text{eg}}^\top$, where $\Gamma_{\text{eg}} = (u_{\text{eg},1}, \cdots, u_{\text{eg},n}) \in \mathbb{R}^{n\times n}$ are eigenvectors and $\Psi = \text{diag}(\psi_1, \cdots, \psi_n)$ is the diagonal matrix of the associated eigenvalues in a deceasing order. Denote $\bar{\zeta}_{\boldsymbol{\beta}} = \zeta_{\boldsymbol{\beta}}/\|\zeta_{\boldsymbol{\beta}}\| = \mathbf{X}\boldsymbol{\beta}_{Q_x}/\|\mathbf{X}\boldsymbol{\beta}_{Q_x}\|$. When an initial estimator $\hat{\boldsymbol{\beta}}$ is available, $\bar{\zeta}_{\boldsymbol{\beta}}$ can be estimated by $\bar{\zeta}_{\hat{\boldsymbol{\beta}}}$. To utilize two different sources of information jointly, or in other words to use both $\bar{\zeta}_{\hat{\boldsymbol{\beta}}}$ and the columns of Γ_{eg}, we construct Γ by replacing one of the columns (say for example the ith column) of Γ_{eg} by $\bar{\zeta}_{\hat{\boldsymbol{\beta}}}$, that is, $\Gamma = \Gamma(\hat{\boldsymbol{\beta}})$, defined as

$$\Gamma(\hat{\boldsymbol{\beta}}) = (\bar{\zeta}_{\hat{\boldsymbol{\beta}}}, u_{\text{eg},j}, j \neq i), \tag{2.5}$$

which is the empirical version of $\Gamma(\boldsymbol{\beta}) = (\bar{\zeta}_{\boldsymbol{\beta}}, u_{\text{eg},j}, j \neq i)$. More discussion on this process is provided in Proposition 1 and Remark 2 below.

Recall that α_x, associated with predictor $\sqrt{n} z_x$, is the parameter of interest in the transformed model that has predictors $\sqrt{n}(z_x, \Gamma)$. It is desirable to avoid the collinearity between z_x and other predictors in the transformed model. Hence, we assume that $\hat{\boldsymbol{\beta}}$ satisfies that $\bar{\zeta}_{\hat{\boldsymbol{\beta}}} \neq z_x$, which can be checked from data, and require the matrix $(z_x, u_{\text{eg},j}, j \neq i)$ being invertible. Note that it is also required that $\Gamma(\hat{\boldsymbol{\beta}})$, i.e. $(\bar{\zeta}_{\hat{\boldsymbol{\beta}}}, u_{\text{eg},j}, j \neq i)$, is invertible.

Proposition 1. *Suppose that $\hat{\boldsymbol{\beta}}$ satisfies $|z_x^\top \bar{\zeta}_{\hat{\boldsymbol{\beta}}}| > 0$ and $\bar{\zeta}_{\hat{\boldsymbol{\beta}}} \neq z_x$. Then for at least one $i \in \{1, \cdots, n\}$, it holds that both matrices $(\bar{\zeta}_{\hat{\boldsymbol{\beta}}}, u_{\text{eg},j}, j \neq i)$ and $(z_x, u_{\text{eg},j}, j \neq i)$ are invertible or equivalently that* $\min\{|u_{\text{eg},i}^\top \bar{\zeta}_{\hat{\boldsymbol{\beta}}}|, |u_{\text{eg},i}^\top z_x|\} > 0$.

In principle, we can replace any $u_{\text{eg},i}$ by $\bar{\zeta}_{\hat{\boldsymbol{\beta}}}$ as long as $\min\{|u_{\text{eg},i}^\top \bar{\zeta}_{\hat{\boldsymbol{\beta}}}|, |u_{\text{eg},i}^\top z_x|\} > 0$. In our numerical studies, the strategy in Remark 2 below is used to further reduce collinearity.

Remark 2. *Reducing the collinearity between z_x and other predictors in the transformed model makes the estimator of α_x more stable. In our simulation studies, we replace u_{eg,i_0} by $\bar{\zeta}_{\hat{\boldsymbol{\beta}}}$ with $i_0 = arg \max_{1 \leq i \leq n} |u_{\text{eg},i}^\top z_x|$, and the resulting $\Gamma(\hat{\boldsymbol{\beta}})$ is observed nonsingular numerically.*

Naturally, one can just use the information in Source II by setting $\Gamma = \Gamma_{\text{eg}}$. A good property for this choice is that it depends only on $\mathbf{X}$ without requiring an initial estimator $\hat{\boldsymbol{\beta}}$ and is free of any assumption on the sparsity of $\boldsymbol{\beta}$. However, this choice may not be ideal if a reasonably good initial estimator $\hat{\boldsymbol{\beta}}$ can be obtained. We summarize the constructions of Γ used in this paper in Table 1 below.

Table 1: Candidates of Γ for the pointwise estimator

1. $\Gamma = \Gamma_{\text{eg}}$, where $\Gamma_{\text{eg}} \in \mathbb{R}^{n\times n}$ contains the eigenvectors of $n^{-1}\mathbf{X}Q_x\mathbf{X}^\top$.
2. $\Gamma = \Gamma(\hat{\boldsymbol{\beta}}) = (\bar{\zeta}_{\hat{\boldsymbol{\beta}}}, u_{\text{eg},j}, j \neq i_0)$, where i_0 can be taken as any i satisfying the condition in Proposition 1 or selected by the procedure in Remark 2, when an initial estimator $\hat{\boldsymbol{\beta}}$ is available.

In Section 3, we show that when $\boldsymbol{\lambda}(n^{-1}\mathbf{X}Q_x\mathbf{X}^\top)$ is sparse enough, $\boldsymbol{\theta} = \Gamma_{\text{eg}}^{-1}\zeta_{\boldsymbol{\beta}}$ is approximately sparse without any sparsity assumption on $\boldsymbol{\beta}$. An extreme case is Σ being a low rank matrix, where $\boldsymbol{\theta}$ is exactly sparse with at most $rank(\Sigma) + 1$ nonzero elements. It is worth pointing out that which elements of $\boldsymbol{\theta}$ are large are generally unknown since $\boldsymbol{\beta}$ is involved. When $\boldsymbol{\lambda}(n^{-1}\mathbf{X}Q_x\mathbf{X}^\top)$ is less sparse, as discussed below, $\hat{\boldsymbol{\beta}}$ will provide additional information and $\boldsymbol{\theta}$ can still be approximately sparse.

When $\Gamma = \Gamma(\hat{\boldsymbol{\beta}})$, the sparsity of $\boldsymbol{\theta}$ also depends on the accuracy of $\hat{\boldsymbol{\beta}}$. For the ideal case that $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, it can be shown that $\boldsymbol{\theta} = \Gamma(\boldsymbol{\beta})^{-1}\zeta_{\boldsymbol{\beta}} \propto e_1 = (1, 0, \cdots, 0)^\top$, and consequently $\boldsymbol{\theta}$ is a sparse vector. This argument still holds, if we replace $(u_{\text{eg},j}, j \neq i_0)$ by any other vectors such that $\Gamma(\boldsymbol{\beta})$ is invertible, implying that if we know $\boldsymbol{\beta}$, there is no need for additional information. Consequently, if $\hat{\boldsymbol{\beta}}$ is good, $(u_{\text{eg},j}, j \neq i_0)$ do not help much. When $\hat{\boldsymbol{\beta}}$ is not good enough (e.g. $\boldsymbol{\beta}$ is less sparse in particular) but $\boldsymbol{\lambda}(n^{-1}\mathbf{X}Q_x\mathbf{X}^\top)$ is sufficiently sparse, using $u_{\text{eg},i}$'s will compensate the low accuracy of $\hat{\boldsymbol{\beta}}$. Thus both types of information can help each other in our framework and the estimator becomes more robust to the underlying assumptions. Details are provided in Section 3.

2.3 Penalized estimator and pointwise estimator

As discussed in Section 2.2, the parameters in the transformed model (2.3) can be approximately sparse if Γ is constructed properly. Thus we consider the minimization of the following objective function

$$L_{\lambda,\Gamma}(\boldsymbol{\alpha}) = n^{-1}\|\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha}\|^2 + P_{\text{en},\lambda}(\boldsymbol{\alpha}), \tag{2.6}$$

where $P_{\text{en},\lambda}(\boldsymbol{\alpha}) = \lambda\|\boldsymbol{\alpha}\|_1$ is the ℓ_1 penalty function used by LASSO, and λ and Γ are the tuning parameters. The parameter λ plays the same role as that for the usual regularized estimator and can be selected by cross validation (CV). The selection of Γ will be elaborated below. Since the ℓ_1 penalty usually induces biases for coefficients of large absolute values, to solve this problem, other nonconvex penalty functions, such as the SCAD (Fan and Li, 2001) or MCP (Zhang, 2010), can be used instead. Denote the minimizer as

$$\hat{\boldsymbol{\alpha}}_{\lambda,\Gamma} = (\hat{\alpha}_x, \hat{\boldsymbol{\theta}}^\top)^\top = \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^{n+1}} L_{\lambda,\Gamma}(\boldsymbol{\alpha}). \tag{2.7}$$

Then we have $\hat{\gamma}_x = \hat{\alpha}_x \cdot \|x\|^2\|\mathbf{X}x\|^{-1}n^{1/2}$.

Remark 3. *Our approach involves eigenvalue decomposition of the matrix $n^{-1}\mathbf{X}Q_x\mathbf{X}^\top \in \mathbb{R}^{n\times n}$ with the computational complexity of the order $O(n^3)$, which can be a burden when n is large. To reduce the complexity, the Divide-and-Conquer (DC) approach for handling big data can be used (Zhang et al., 2015). Simulation results based on DC are presented in Section G.2 of the Supplementary Materials.*

Next we briefly discuss the estimator in (2.7). First, the predictor $\mathbf{Z}$ in Model (2.3) involves the transformation matrix Γ, while in the classical model (2.1), the predictor is $\mathbf{X}$, of which each row represents a realization of the predictors. Second, unlike the classical methods such as the LASSO, where the parameters are unknown constants, the parameter $\boldsymbol{\alpha}$ here involves $(x, \mathbf{X}, \Gamma)$, which makes the theoretical analysis challenging.

Remark 4. *In the above arguments, we consider a single point x that may or may not be from the same distribution as the predictor vector. However, if we are going to consider a large number of test points $\{x_i, i = 1, \cdots, n_{te}\}$ that are i.i.d. observations of X_i, the bias should be taken into account. Because $E(\gamma_{x_i}) = E(\boldsymbol{\beta}^\top x_i) = 0$ and $\text{var}(\gamma_{x_i}) = \boldsymbol{\beta}^\top\Sigma\boldsymbol{\beta} <$*

$\mathrm{var}(Y_1)$*, there will be many* γ_{x_i}*'s that are close to 0 and the corresponding estimators are shrunken to 0 by the penalization, resulting in large biases in terms of the average estimation error. There are many bias correction methods for LASSO and related penalized estimators in the literature (Belloni and Chernozhukov, 2013; Zhang and Zhang, 2014, etc.). In our simulation results for the prediction problem, the method of Belloni and Chernozhukov (2013) is applied.*

For clarity, we briefly summarize the estimation procedure for a given Γ as follows.

Algorithm 1: Estimator of γ_x (or $\gamma_{\tilde{x}_{S_1}}$) with given Γ

1. (Estimate α_x) Solve the optimization problem (2.6), where the optimal λ is chosen by CV; the corresponding parameter obtained is denoted as $(\hat{\alpha}_x, \hat{\boldsymbol{\theta}}^\top)^\top$.
2. (Bias-correction) This is an optional step. Denote $\hat{S} = \mathrm{supp}(\hat{\boldsymbol{\theta}})$ and $\Gamma_{\cdot\hat{S}}$ the columns of Γ with index $\hat{S}$. Apply the OLS with responses $\mathbf{Y}$ and predictors $\mathbf{Z} = \sqrt{n}(z_x, \Gamma_{\cdot\hat{S}})$ to get the updated coefficient $\hat{\alpha}_x$ of $\sqrt{n}z_x$, which is a bias-corrected estimator of α_x.
3. (Estimate γ_x) γ_x is estimated by $\hat{\gamma}_x = \hat{\alpha}_x \cdot \|x\|^2 \|\mathbf{X}x\|^{-1} n^{1/2}$.
4. (Alternative Steps 1–3) Replacing x by $\tilde{x}_{S_1}$ in Steps 1–3, we get the estimator $\hat{\gamma}_{\tilde{x}_{S_1}}$.

Step 2 is mainly applied for the setting in Remark 4 with a large number of testing points that are i.i.d. observations as of the training data X_i's. For clarity, the pointwise estimator obtained by Algorithm 1 is named based on the specific Γ used. There are many estimators of $\boldsymbol{\beta}$ available for different settings in the literature, and can be used as an initial estimator. Denote by $\hat{\boldsymbol{\beta}}_{\text{lasso}}$ and $\hat{\boldsymbol{\beta}}_{\text{ridge}}$ the estimators of LASSO and ridge regression respectively, and by $\hat{\boldsymbol{\beta}}_{\text{rdl}}$ the overparameterized ridgeless OLS estimator using the Moore-Penrose generalized inverse (Bartlett et al., 2020; Azriel and Schwartzman, 2020; Hastie et al., 2022). We consider the estimators $\hat{\gamma}_x$ in Algorithm 1 with Γ being Γ_{eg} and $\Gamma(\hat{\boldsymbol{\beta}})$ for $\hat{\boldsymbol{\beta}} \in \{\hat{\boldsymbol{\beta}}_{\text{lasso}}, \hat{\boldsymbol{\beta}}_{\text{ridge}}, \hat{\boldsymbol{\beta}}_{\text{rdl}}\}$, and the resulting estimators are denoted as $P_{\text{eg}}, P_{\text{lasso}}, P_{\text{ridge}}, P_{\text{rdl}}$ respectively. Now we propose a procedure to select Γ adaptively. Denote by $\mathcal{M}$ a set of estimators, e.g. $\mathcal{M} = \{P_{\text{lasso}}, P_{\text{ridge}}, P_{\text{eg}}, P_{\text{rdl}}\}$ in Setting 1 of our numerical study. The best estimator selected from $\mathcal{M}$ by the following CV procedure will be denoted as PWE.

Algorithm 2: Select Γ by CV

1. Compute the CV error for estimators in $\mathcal{M}$. Split randomly the whole data $\mathcal{D}$ of size n into K parts, denoted as $D_1, \cdots, D_K$. For each method $A \in \mathcal{M}$, compute the CV error, $n^{-1} \sum_{k=1}^{K} \sum_{(x_i, y_i) \in D_k} (\hat{\gamma}_{x_i}^A - y_i)^2$, where $\hat{\gamma}_{x_i}^A$ is estimated by A with data $\mathcal{D} \setminus D_k$.
2. The method A_0 in $\mathcal{M}$ with the minimum CV error is chosen to be the best one.
3. The final estimator of γ_x is $\hat{\gamma}_x^{A_0}$.

For Example 1 in Section 2.1 where x is a sparse vector, we apply Algorithm 2 to get the pointwise estimator. For the prediction problems in Example 2, since the sparsity degree of $\boldsymbol{\beta}$ is unknown, we have two choices on the subset S_1: (1) simply taking $S_1 = \{1, \cdots, p\}$ (i.e. estimate γ_x directly); (2) estimating S_1 from data using LASSO or screening methods. Given Γ, among the candidates of S_1, we can select the best one by CV, similar to Steps 1 and 2 of Algorithm 2. Note that x_i should be replaced by $\tilde{x}_{iS_1}$, which is defined in the way similar to $\tilde{x}_{S_1}$, when one computes the CV error in Step 1 of Algorithm 2. Moreover, one can select both Γ and S_1 simultaneously by CV.

3 Properties of the penalized estimator $\hat{\alpha}_x$

Throughout our theoretical analysis, it is assumed that ϵ_i's are *i.i.d.* from $N(0, \sigma^2)$, and $(x, \mathbf{X})$ can be fixed or random and will be specified later. In this section, we present theoretical properties of the regularized estimator $\hat{\alpha}_x$ in (2.7), which lays a foundation for properties of the estimator $\hat{\gamma}_x$ in Section 4. To this end, we first establish results for a generic invertible matrix Γ with fixed $(x, \mathbf{X})$ in Section 3.1, and then apply it to Γ_{eg} and $\Gamma(\hat{\boldsymbol{\beta}})$ in Sections 3.2 and 3.3, respectively.

3.1 A general result on $\hat{\alpha}_x$ with a generic Γ and fixed $(x, \mathbf{X})$

Recall that $\mathbf{Z} = \sqrt{n}(z_x, \Gamma)$ in Model (2.3). Let $v_x = (1, -b_x^\top)^\top/(1+b_x^\top b_x)^{1/2}$ with $b_x = \Gamma^{-1}z_x$. It can be shown that v_x is the eigenvector associated with the zero eigenvalue of $n^{-1}\mathbf{Z}^\top\mathbf{Z}$ (see proof of Theorem E.1 in Supplementary Materials). For any $\boldsymbol{v} = (v_1, \cdots, v_{n+1})^\top \in \mathbb{R}^{n+1}$, denote $\|\boldsymbol{v}\|_{(q)} = \sum_{i=1}^{n+1} |v_i|^q; q \in [0, 1]$. It is worth noting that $\|\boldsymbol{v}\|_{(q)}$ here is not the ℓ_q norm of $\boldsymbol{v}$ that is defined as $\|\boldsymbol{v}\|_q = \|\boldsymbol{v}\|_{(q)}^{1/q}$; and for $q = 1$, they are the same. This notation $\|\boldsymbol{v}\|_{(q)}$ is convenient for our purpose, particularly for the discussions of the case with $q = 0$ or $q \to 0$.

For $\boldsymbol{\alpha} = (\alpha_x, \boldsymbol{\theta}^\top)^\top$ and $q \in [0, 1]$, denote $R_q = \|\boldsymbol{\alpha}\|_{(q)} = |\alpha_x|^q + \|\boldsymbol{\theta}\|_{(q)}$, which measures the sparsity of the parameter $\boldsymbol{\alpha}$. Particularly, when $q = 0$, R_q is the number of nonzero elements in $\boldsymbol{\alpha}$. Let the tuning parameter $\lambda = \lambda_n := a_0\sigma\sqrt{(\log n)/n}$ with $a_0 \geq 2$. We have the following result on a generic invertible matrix Γ.

Theorem 1. *Assume that $(x, \mathbf{X})$ are fixed. Let Γ be a generic invertible matrix that may depend on $(x, \mathbf{X})$. Assume the following $\boldsymbol{\alpha}$-sparsity condition: $R_q \leq C\lambda_n^q \|v_x\|_1^2$ for some $q \in [0, 1]$ and some constant $C > 0$. Then with probability $1 - C_1 n^{-3}$, we have*

$$|\hat{\alpha}_x - \alpha_x| \leq \frac{5}{4}\lambda_n + C_\kappa \lambda_n^{1-q} R_q \cdot \min\{\lambda_n^{q/2} R_q^{-1/2} \|\Gamma^\top z_x\|, \|\Gamma^\top z_x\|_\infty\},$$

where C_1, C_κ are positive constants. Furthermore, when Γ is a generic orthogonal matrix, the $\boldsymbol{\alpha}$-sparsity condition can be simplified as $R_q \leq C'\lambda_n^q \|\Gamma^\top z_x\|_1^2$ for some constant $C' > 0$.

We make a brief discussion on the above result. First, note that $\|\Gamma^\top z_x\|_\infty \leq \|\Gamma^\top z_x\|$, where the latter equals 1 when Γ is an orthogonal matrix. Further discussions are referred to Section A of the Supplementary Materials. Second, since λ_n has the order $\sqrt{\log n/n}$, the bound of Theorem 1 does not explicitly depend on p, which is reasonable, since we have only $n + 1$ parameters in the transformed model (2.3). Third, the bound given above involves the sparsity of the parameter vector $\boldsymbol{\alpha}$ in the transformed model instead of that of $\boldsymbol{\beta}$, allowing $\boldsymbol{\beta}$ being sparse or non-sparse. As an application of Theorem 1, we present the results for Γ being Γ_{eg} and $\Gamma(\hat{\boldsymbol{\beta}})$ respectively below.

Proposition 2. *Suppose that* $(x, \mathbf{X})$ *are fixed. Let* $\Gamma = \Gamma_{\text{eg}}$*. Assume that the* $\boldsymbol{\alpha}$*-sparsity condition* $R_q \leq C\lambda_n^q \|\Gamma^\top z_x\|_1^2$ *holds for some constant* $C > 0$*. Then it holds that* $|\hat{\alpha}_x - \alpha_x| \leq C' \lambda_n^{1-q/2} R_q^{1/2}$ *with probability* $1 - C_1 n^{-3}$*, where* C_1, C' *are positive constants.*

Similar to Theorem 1, Proposition 2 allows $\boldsymbol{\beta}$ to be sparse or non-sparse. Due to α_x being bounded, we have $R_q \asymp \|\boldsymbol{\theta}\|_{(q)}$, which can be bounded as shown in Section 3.2. When X_i's are random, a lower bound on $\|\Gamma_{\text{eg}}^\top z_x\|_1$ is given in Section E of the Supplement.

3.2 Sparsity of $\boldsymbol{\theta}$ and properties of $\hat{\alpha}_x$ when $\Gamma = \Gamma_{\text{eg}}$

We first show that for $\Gamma = \Gamma_{\text{eg}}$, if the eigenvalues of $n^{-1}\mathbf{X}Q_x\mathbf{X}^\top$ decrease at a certain rate, $\boldsymbol{\theta}$ will be approximately sparse; specifically $\|\boldsymbol{\theta}\|_{(q)}$ is bounded for some $q \in [0, 1]$. Then we give the asymptotic results on $\hat{\alpha}_x$. Recall $\psi_1 \geq \psi_2 \cdots \geq \psi_n \geq 0$ are the eigenvalues of $n^{-1}\mathbf{X}Q_x\mathbf{X}$; they are also the first n eigenvalues of $n^{-1}Q_x\mathbf{X}^\top\mathbf{X}Q_x$. For $q \in [0, 1]$ and $k = 0, 1, \cdots, n-1$, let $\bar{\psi}_{q,k} = (n-k)^{-1} \sum_{i=k+1}^{n} \psi_i^{q/(2-q)}$ and $\phi_k^{1/2}(\boldsymbol{\beta})$ be the norm of the projected vector of $\boldsymbol{\beta}$ onto the subspace spanned by the eigenvectors of $n^{-1}Q_x\mathbf{X}^\top\mathbf{X}Q_x$ associated with eigenvalues $\psi_{k+1}, \cdots, \psi_n$, $\phi_n(\boldsymbol{\beta}) = \bar{\psi}_{q,n} = 0$, and $\bar{\psi}_{q,k}$ is the average magnitude of the smallest $n-k$ eigenvalues. For any $x_1, x_2 \geq 0$, denote $H_{q,k}(x_1, x_2) = k^{1-q/2} + [(n-k)x_1]^{1-q/2} x_2^{q/2}$. The following Lemma 1 is a deterministic result on $\|\boldsymbol{\theta}\|_{(q)}$.

Lemma 1. *For* $q \in [0, 1]$*, it holds that* $\|\boldsymbol{\theta}\|_{(q)} \leq \min_{0 \leq k \leq n} \left\{ \|\zeta_{\boldsymbol{\beta}}\|^q k^{1-q/2} + \left[(n-k)\bar{\psi}_{q,k}\right]^{1-q/2} \phi_k^{q/2}(\boldsymbol{\beta}) \right\}$*. Moreover, since* $\|\zeta_{\boldsymbol{\beta}}\| = O(1)$ *in Model (2.3), it holds that* $\|\boldsymbol{\theta}\|_{(q)} = O\Big(\min_{0 \leq k \leq n} H_{q,k}(\bar{\psi}_{q,k}, \phi_k(\boldsymbol{\beta})) \Big)$*.*

Generally, it is difficult to obtain the sharp upper bound in Lemma 1 without any information of eigenvalues and $\boldsymbol{\beta}$. Simply setting $k = n$, we have the *trivial bound*

$$\|\boldsymbol{\theta}\|_{(q)} \lesssim H_{q,n}(\bar{\psi}_{q,n}, \phi_k(\boldsymbol{\beta})) = n^{1-q/2}, \tag{3.1}$$

which however is unbounded and undesirable. To get better bounds on $\|\boldsymbol{\theta}\|_{(q)}$, we take into account of the properties of eigenvalues and $\boldsymbol{\beta}$, and consider the following three cases:

Case (a) : $n^{-1}\mathbf{X}\mathbf{X}^\top$ has low rank, say $rank(n^{-1}\mathbf{X}\mathbf{X}^\top) = r_{\mathbf{X}} = O(1)$;
Case (b) : $\boldsymbol{\beta}$ is sparse with $\|\boldsymbol{\beta}\|_0 = s_0$ and that $\|\boldsymbol{\beta}\|_\infty < C < \infty$;
Case (c) : $\boldsymbol{\beta}$ is dense, following a normal distribution $N(0, \Sigma_{\boldsymbol{\beta}})$.

Results are presented in Corollary 1 and Proposition 3, respectively.

Corollary 1. *The following conclusions are deterministic.*

(1) For Case (a), it holds that $\bar{\psi}_{q,k} = 0$ *for* $k \geq r_{\mathbf{X}} + 1$, *and* $\|\boldsymbol{\theta}\|_{(q)} = O(r_{\mathbf{X}}^{1-q/2}) = O(1)$.

(2) For Case (b), assume that $n < p$, *then* $\|\boldsymbol{\theta}\|_{(q)} \lesssim \min_{0 \leq k \leq n} H_{q,k}(\bar{\psi}_{q,k}, s_0)$. *Thus if* $\bar{\psi}_{q,k_0} = O(s_0^{-q/(2-q)} n^{-1})$ *for some fixed* k_0, *then* $\|\boldsymbol{\theta}\|_{(q)} = O(k_0^{1-q/2}) = O(1)$.

In Corollary 1, conclusion (1) holds regardless of the sparsity of $\boldsymbol{\beta}$, while conclusion (2) relaxes the conditions on eigenvalues by taking advantages of the sparsity of $\boldsymbol{\beta}$. For Case (c), we assume that the column space of $\Sigma_{\boldsymbol{\beta}}$ is the same as that of Σ, accommodating the identifiability condition that $\boldsymbol{\beta} \in \text{span}(\Sigma)$. We have the following results.

Proposition 3. *Suppose that* $(\mathbf{X}, x)$ *is fixed. For Case (c) with span(*$\Sigma_{\boldsymbol{\beta}}$*)=span(*$\Sigma$*), assume the following conditions: (i)* $\|\zeta_{\boldsymbol{\beta}}\| = O_p(1)$; *(ii)* $(np)^{-1} \sum_{i=1}^{n} X_i^\top Q_x X_i \asymp 1$; *(iii)* $n < p$. *Then it holds that* $\phi_k(\boldsymbol{\beta}) = O_p(d_{\boldsymbol{\beta}} n/p)$ *for all* k *and that* $\|\boldsymbol{\theta}\|_{(q)} = O_p\Big(\min_{0 \leq k \leq n} H_{q,k}\left(\bar{\psi}_{q,k}, d_{\boldsymbol{\beta}} n/p\right)\Big)$, *where* $d_{\boldsymbol{\beta}} = \lambda_{\max}(\Sigma_{\boldsymbol{\beta}})/\lambda_{\min}^{+}(\Sigma_{\boldsymbol{\beta}})$ *can be viewed as the condition number of* $\Sigma_{\boldsymbol{\beta}}$. *For clarity, we assume that* $d_{\boldsymbol{\beta}} = O(1)$ *and consider the following two specific examples of eigenvalues:*

(1) Suppose that $\psi_1 \geq \cdots \geq \psi_{k_0} = \Omega(p)$ *and* $\max_{j \geq k_0+1} \psi_j = O(pn^{-2/q})$ *for some fixed* k_0. *Then it holds that* $\|\boldsymbol{\theta}\|_{(q)} = O_p(k_0^{1-q/2}) = O_p(1)$.

(2) Denote $f_\psi(i) = \psi_i / \sum_{i=1}^{n} \psi_i$, *the scaled version of* ψ_i, $i = 1, \cdots, n$. *Assume that* $f_\psi(i)$ *decreases exponentially, that is,* $f_\psi(w) = a\exp(-a(w-1))$ *for* $w \geq 1$, *where* $a = a_n > 0$ *may depend on* n. *Then* $\|\boldsymbol{\theta}\|_{(q)} = O_p(1)$ *if* $a_n = \Omega(q^{-1}\log n)$.

Condition (i) is a natural extension of the condition $\|\zeta_{\boldsymbol{\beta}}\| = O(1)$ for fixed $\boldsymbol{\beta}$ in Section 2. Condition (ii) is mild, ruling out the extreme case that Σ has eigenvalues $(p, 0, \cdots, 0)$. Details are referred to the proof in Supplementary Materials. The two examples in Proposition 3 imply that $\boldsymbol{\theta}$ can be approximately sparse, when eigenvalues decrease fast enough. Based on the results of $\|\boldsymbol{\theta}\|_{(q)}$ in Corollary 1 and Propositions 3, by applying the conclusion of Proposition 2, we are ready to give the asymptotic results of $\hat{\alpha}_x$.

Theorem 2. *Suppose that $(x, \mathbf{X})$ are fixed. Taking $\Gamma = \Gamma_{\mathrm{eg}}$, we have the following conclusions for Cases (a)-(c):*

(1) For Case (a), $|\hat{\alpha}_x - \alpha_x| = O_p\left(\lambda_n r_{\mathbf{X}}^{1/2}\right)$. For Case (b), suppose that $\lambda_n^q \|\Gamma^\top z_x\|_1^2 = \Omega(1)$ and that $n < p$; if $\bar{\psi}_{q,k_0} = O(s_0^{-q/(2-q)} n^{-1})$ for some fixed k_0, then $|\hat{\alpha}_x - \alpha_x| = O_p(\lambda_n^{1-q/2} k_0^{1/2-q/4})$.

(2) For Case (c) of a dense $\boldsymbol{\beta}$, assume that the conditions of Proposition 3 hold and that $\lambda_n^q \|\Gamma^\top z_x\|_1^2 = \Omega(1)$. If ψ_i's are from Proposition 3 (1), then $|\hat{\alpha}_x - \alpha_x| = O_p(\lambda_n^{1-q/2} k_0^{1/2-q/4})$; if ψ_i's are from Proposition 3 (2), then $|\hat{\alpha}_x - \alpha_x| = O_p(\lambda_n^{1-q/2})$.

The condition $\lambda_n^q \|\Gamma^\top z_x\|_1^2 = \Omega(1)$ above can be checked from data. Here k_0 and $r_{\mathbf{X}}$ reflect the sparsity of the parameters in the transformed model. Faster decay rates of eigenvalues result in smaller values of k_0 and $r_{\mathbf{X}}$, and consequently better convergence rates of $|\hat{\alpha}_x - \alpha_x|$. Moreover, a smaller value of q results in a smaller value of $\lambda_n^{1-q/2}$, but requires a faster decreasing rate of eigenvalues. Note that the above bounds only depend on n and the degree of sparsity in eigenvalues, and do not explicitly depend on p. Moreover, for Cases (a) and (c) where $\boldsymbol{\beta}$ is allowed to be dense, the sparsity of eigenvalues is complementary to the sparsity of $\boldsymbol{\beta}$.

3.3 Properties of the regularized estimator $\hat{\alpha}_x$ with an initial $\hat{\boldsymbol{\beta}}$

We study properties of our estimator with $\Gamma(\hat{\boldsymbol{\beta}})$, giving the convergence rate of $\hat{\alpha}_x$, for fixed x and X_i's being *i.i.d.* from $N(0, \Sigma)$. The normality assumption of X_i's simplifies the proofs and can be relaxed to general distributions such as sub-Gaussian distributions. We first give a result on $\|\bar{\zeta}_{\hat{\boldsymbol{\beta}}} - \bar{\zeta}_{\boldsymbol{\beta}}\|$. Recall that $\zeta_{\hat{\boldsymbol{\beta}}} = n^{-1/2}\mathbf{X}\hat{\boldsymbol{\beta}}_{Q_x}$.

Proposition 4. *Assume that x is fixed, X_i's are i.i.d. from $N(0, \Sigma)$, and $\mathrm{cov}(X_i^\top \boldsymbol{\beta}_{Q_x}) \asymp 1$. For any estimator $\hat{\boldsymbol{\beta}}$ in Model (2.1), it holds that $\|\bar{\zeta}_{\hat{\boldsymbol{\beta}}} - \bar{\zeta}_{\boldsymbol{\beta}}\| = O_p(\min\{2, \|\zeta_{\hat{\boldsymbol{\beta}}} - \zeta_{\boldsymbol{\beta}}\|\})$. Assuming further that $\|x\|_\infty \|x\|_\Sigma / \|x\|^2 = O(1)$, then $\|\zeta_{\hat{\boldsymbol{\beta}}} - \zeta_{\boldsymbol{\beta}}\| = O_p(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 (\log p)^{1/2})$.*

When $\hat{\boldsymbol{\beta}}$ is the LASSO estimator, we have $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 = O_p(\sigma s_0 \sqrt{\log p/n})$ with $s_0 = \mathrm{supp}(\boldsymbol{\beta})$ (Bickel et al., 2009), and consequently $\|\bar{\zeta}_{\hat{\boldsymbol{\beta}}} - \bar{\zeta}_{\boldsymbol{\beta}}\| = O_p(\min\{2, \sigma s_0 \sqrt{(\log p)^2/n}\})$. Let $H_{1,k}(\bar{\psi}_{1,k}, x)$ be the function obtained by setting $q = 1$ in $H_{q,k}(\bar{\psi}_{q,k}, x)$ defined in Section 3.2.

Theorem 3. *Let $\Gamma = \Gamma(\hat{\boldsymbol{\beta}})$. Assume that following conditions: (i) X_i's are i.i.d. from $N(0, \Sigma)$, x is fixed, and $n < p$; (ii) $\boldsymbol{\beta}$ satisfies $\mathrm{cov}(X_i^\top \boldsymbol{\beta}_{Q_x}) \asymp 1$ and $\hat{\boldsymbol{\beta}}$ is obtained from additional data independent of $(\mathbf{X}, \mathbf{Y})$, satisfying $\mathrm{cov}(X_i^\top \hat{\boldsymbol{\beta}}_{Q_x} | \hat{\boldsymbol{\beta}}) \asymp 1$. Then it holds that $|\hat{\alpha}_x - \alpha_x| = O_p(\lambda_n [1 + H_{\min} \|\zeta_{\hat{\boldsymbol{\beta}}} - \zeta_{\boldsymbol{\beta}}\|])$, where $H_{\min} = \min_{0 \le k \le n} H_{1,k}(\bar{\psi}_{1,k}, n/[p\lambda^+_{\min}(\Sigma)])$. Assuming further the* Complementary condition: *$H_{\min} \|\zeta_{\hat{\boldsymbol{\beta}}} - \zeta_{\boldsymbol{\beta}}\| = O_p(1)$, it follows that $|\hat{\alpha}_x - \alpha_x| = O_p(\lambda_n)$.*

The assumption on $\boldsymbol{\beta}$ in *(ii)* is mild as $\mathrm{cov}(X_i^\top \boldsymbol{\beta}_{Q_x}) = E(\|\zeta_{\boldsymbol{\beta}}\|^2)$ and $\|\zeta_{\boldsymbol{\beta}}\|$ is bounded in probability. We give some examples on the magnitude of $H_{\min}$.

Corollary 2. *(1) If Σ is of low rank, say $rank(\Sigma) = r_\Sigma = O(1)$, then $H_{\min} = O(r_\Sigma^{1/2})$. (2) Suppose that $\max_{j \ge k_0+1} \psi_j = O(pn^{-2})$ for some fixed k_0 and that $\lambda^+_{\min}(\Sigma) \gtrsim 1$. Then $H_{\min} = O(k_0^{1/2})$. (3) Suppose that $\lambda^+_{\min}(\Sigma) \gtrsim 1$ and denote $f_\psi(i) = \psi_i / \sum_{i=1}^n \psi_i$, $i = 1, \cdots, n$.*

Assuming that $f_\psi(i)$ decreases exponentially, i.e. $f_\psi(w) = a\exp(-a(w-1))$ for $w \geq 1$, where $a = a_n \gtrsim \log n$, then $H_{\min} = O(1)$.

Proof of Corollary 2 is similar to those of Corollary 1 and Proposition 3 and is omitted. We point out two basic facts: $1 \lesssim H_{\min} \lesssim \sqrt{n}$ from (3.1), and $\|\zeta_{\hat{\boldsymbol{\beta}}} - \zeta_{\boldsymbol{\beta}}\| = O_p(1)$ by Condition (ii) and the law of large numbers. Moreover, the error rate of an estimator is generally believed having an order no less than $n^{-1/2}$; thus without loss of generality we have $\|\zeta_{\hat{\boldsymbol{\beta}}} - \zeta_{\boldsymbol{\beta}}\| = \Omega_p(n^{-1/2})$ throughout this paper. Hence Theorem 3 leads to the error rate of order $\lambda_n H_{\min}\|\zeta_{\hat{\boldsymbol{\beta}}} - \zeta_{\boldsymbol{\beta}}\| \lesssim \min\{\lambda_n\sqrt{n}\|\zeta_{\hat{\boldsymbol{\beta}}} - \zeta_{\boldsymbol{\beta}}\|, \lambda_n H_{\min}\}$ without any restriction on the decay rate of eigenvalues.

The complementary condition involves two terms, $H_{\min}$ and $\|\zeta_{\hat{\boldsymbol{\beta}}} - \zeta_{\boldsymbol{\beta}}\|$, where the former is controlled by the decay rate of the eigenvalues, and the latter depends on the accuracy of $\hat{\boldsymbol{\beta}}$. As argued before, as long as the eigenvalues decrease fast, $H_{\min}$ will be bounded, even if $\zeta_{\hat{\boldsymbol{\beta}}} - \zeta_{\boldsymbol{\beta}} \nrightarrow 0$. Therefore, if $\hat{\boldsymbol{\beta}}$ is not accurate but eigenvalues decay fast, we can still get good results. The same argument applies when $\hat{\boldsymbol{\beta}}$ is accurate but the eigenvalues decay slowly. Particularly, if $\hat{\boldsymbol{\beta}}$ is good enough such that $\|\zeta_{\hat{\boldsymbol{\beta}}} - \zeta_{\boldsymbol{\beta}}\| = O_p(n^{-1/2})$, then the requirement on the decaying rate of eigenvalues can be removed completely. Thus the information of $\hat{\boldsymbol{\beta}}$ and eigenvalues are complementary to each other, requiring only the product of $H_{\min}$ and $\|\zeta_{\hat{\boldsymbol{\beta}}} - \zeta_{\boldsymbol{\beta}}\|$ being bounded. Theorem 3 assumes $\hat{\boldsymbol{\beta}}$ being independent of the data $(\mathbf{X}, \mathbf{Y})$. When $\hat{\boldsymbol{\beta}}$ depends on $(\mathbf{X}, \mathbf{Y})$, we get a similar result on $\hat{\alpha}_x$ with some modifications on the Condition (i), which is presented in Section F of the Supplementary Materials.

4 Results of $\hat{\gamma}_x$ for two types of test points

In Section 3, we establish theoretical properties of $\hat{\alpha}_x$. With $\hat{\gamma}_x = \hat{\alpha}_x \cdot \|x\|^2\|\mathbf{X}x\|^{-1}n^{1/2}$, it is natural to get the corresponding results on $\hat{\gamma}_x$ for a generic x. As mentioned in Examples

1 and 2 in Section 2.1, we are interested in two typical settings in particular: (1) x is a sparse vector with $S_x = \mathrm{supp}(x)$ and $s_x = |S_x|$, and (2) x is a random vector as an *i.i.d.* copy of X_i (i.e. the prediction problem). Next we give the theoretical results on $\hat{\gamma}_x$ for these two examples, based on the simple fact that $|\hat{\gamma}_x - \gamma_x| \le |\hat{\alpha}_x - \alpha_x| \cdot \|x\|^2 \|\mathbf{X}x\|^{-1} n^{1/2}$.

4.1 Properties of $\hat{\gamma}_x$ for a sparse x

Proposition 5. *Assume the following conditions: (i) x is a fixed sparse vector satisfying $\|x\|_\infty = O(1)$; (ii) $\lambda_{\min}(n^{-1}\mathbf{X}_{S_x}^\top \mathbf{X}_{S_x}) > c > 0$, where $\mathbf{X}_{S_x}$ is formed by the columns of $\mathbf{X}$ with index S_x. Then we have $\|x\|^2\|\mathbf{X}x\|^{-1}n^{1/2} = O(\|x\|) = O\left(s_x^{1/2}\right)$. Moreover, the following results hold.*

(1) Take $\Gamma = \Gamma_{\mathrm{eg}}$ and assume that the conditions of Theorem 2 hold. For Case (a) in Section 3.2, it holds that $|\hat{\gamma}_x - \gamma_x| = O_p\left(\lambda_n \|x\| r_{\mathbf{X}}^{1/2}\right) = O_p\left(\lambda_n (s_x r_{\mathbf{X}})^{1/2}\right)$; for Cases (b) and (c) in Section 3.2, $|\hat{\gamma}_x - \gamma_x| = O_p\left(\lambda_n^{1-q/2}\|x\|\right) = O_p\left(s_x^{1/2}\lambda_n^{1-q/2}\right)$, where k_0 appeared in Theorem 2 is omitted due to $k_0 = O(1)$.

(2) Let $\Gamma = \Gamma(\hat{\boldsymbol{\beta}})$. Assume further that the conditions of Theorem 3 hold. Then it follows that $|\hat{\gamma}_x - \gamma_x| = O_p\left(s_x^{1/2}\lambda_n H_{\min}\|\zeta_{\hat{\boldsymbol{\beta}}} - \zeta_{\boldsymbol{\beta}}\|\right)$; assume further that the Complementary condition holds, then $|\hat{\gamma}_x - \gamma_x| = O_p\left(\|x\|\sqrt{n^{-1}\log n}\right) = O_p\left(\sqrt{n^{-1}s_x \log n}\right)$.

The condition $\lambda_{\min}(n^{-1}\mathbf{X}_{S_x}^\top \mathbf{X}_{S_x}) > c > 0$ is a type of the restricted eigenvalue condition (Bickel et al., 2009). If X_i's are *i.i.d.* variables, $n^{-1}\mathbf{X}_{S_x}^\top \mathbf{X}_{S_x} \to_p \mathrm{cov}(X_{S_x}) = \Sigma_{S_x S_x}$. Recall that $rank(n^{-1}\mathbf{X}^\top\mathbf{X}) = r_{\mathbf{X}}$ in Case (a) of Theorem 2; then the condition $\lambda_{\min}(n^{-1}\mathbf{X}_{S_x}^\top \mathbf{X}_{S_x}) > c > 0$ implies that $s_x \le r_{\mathbf{X}}$ there.

Remark 5. *We briefly discuss the case of $\Sigma = I_p$ or close to I_p for a sparse vector x. For $\Gamma = \Gamma_{\mathrm{eg}}$, using the trivial bound in (3.1) on R_q and taking $q = 1$ in Proposition*

2, we can see that the error of $|\hat{\alpha}_x - \alpha_x|$ has the order $O_p((\log n)^{1/4})$, and consequently $|\hat{\gamma}_x - \gamma_x| = O_p(\|x\|(\log n)^{1/4})$.

We briefly compare our method with the plug-in estimator using the LASSO estimator $\hat{\boldsymbol{\beta}}$ (named briefly as LASSO). For LASSO, the error varies depending on the direction of x, while results of our estimator depend only on the sparsity degree s_x of x. For simplicity of comparison, we consider a bound of LASSO depending only on s_x. Specifically, as $\|x\|_\infty = O(1)$, we have $|x^\top \hat{\boldsymbol{\beta}}_{\text{lasso}} - x^\top \boldsymbol{\beta}| = O(\|x\| \|\hat{\boldsymbol{\beta}}_{\text{lasso}} - \boldsymbol{\beta}\|) = O_p(\sqrt{s_x s_0 \log p/n})$ (Bickel et al., 2009); the latter will be used as the error rate of LASSO. Recall that P_{eg} denote our estimator with $\Gamma = \Gamma_{\text{eg}}$. For Case (a), it follows from Proposition 5 that P_{eg} is better than LASSO if and only if $r_{\mathbf{X}}(\log n)(\log p)^{-1} = o(s_0)$. Cases (b) and (c) can be analyzed similarly. Recall that P_{lasso} is our estimator with $\Gamma = \Gamma(\hat{\boldsymbol{\beta}}_{\text{lasso}})$.

Corollary 3. *Denote by $T_{n,\text{fix}}$ the ratio of error rate of P_{lasso} over that of LASSO for a fixed sparse x. Suppose that the conditions (i) and (ii) in Proposition 5 and the conditions of Theorem 3 hold. Then P_{lasso} has the error rate of order $\lambda_n H_{\min}\sqrt{s_x(s_0 \log p)^2/n}$ and consequently $T_{n,\text{fix}} = O_p(\lambda_n H_{\min}\sqrt{s_0 \log p})$. If $H_{\min} = o_p(n^{1/2}[s_0(\log n)(\log p)]^{-1/2})$, then $T_{n,\text{fix}} = o_p(1)$, impliying P_{lasso} is superior to LASSO; otherwise P_{lasso} is inferior or similar to LASSO.*

The proof of Corollary 3 is a simple combination of Proposition 4 and (2) of Proposition 5 and is omitted here. Since we always have $H_{\min} \lesssim \sqrt{n}$, the requirement $H_{\min} = o_p(n^{1/2}[s_0(\log n)(\log p)]^{-1/2})$ is mild.

4.2 Properties of $\hat{\gamma}_x$ for prediction problems

In a prediction problem, x and X_i's are *i.i.d.* variables. Recall that $M_\Sigma = \sqrt{tr^2(\Sigma)/tr(\Sigma^2)}$.

Proposition 6. *Suppose that x and X_i's are i.i.d. from $N(0, \Sigma)$. Then $\|x\|^2\|\mathbf{X}x\|^{-1}n^{1/2} = O_p(\|x\|^2/\|x\|_\Sigma) = O_p(M_\Sigma)$. In addition, it holds that $1 \leq M_\Sigma \leq p^{1/2}$. Particularly, $M_\Sigma \asymp 1$ when Σ is of low rank, and $M_\Sigma = p^{1/2}$ when $\Sigma = I_p$.*

Next we first derive properties of $\hat{\gamma}_x$ with $\Gamma = \Gamma_{\text{eg}}$. Then we consider the estimator $\hat{\gamma}_{\tilde{x}_{S_1}}$ with $\Gamma = \Gamma(\hat{\boldsymbol{\beta}})$, given both an initial estimator $\hat{\boldsymbol{\beta}}$ and a subset S_1.

4.2.1 Properties of $\hat{\gamma}_x$ for x in prediction with $\Gamma = \Gamma_{\text{eg}}$

Recall that $|\hat{\gamma}_x - \gamma_x| \leq |\hat{\alpha}_x - \alpha_x| \cdot \|x\|^2\|\mathbf{X}x\|^{-1}n^{1/2}$. Combining Proposition 6 and the result on $|\hat{\alpha}_x - \alpha_x|$ with fixed $(x, \mathbf{X})$ given in Proposition 2 for $\Gamma = \Gamma_{\text{eg}}$, it can be inferred that $|\hat{\gamma}_x - \gamma_x| = O_p\left(\lambda_n^{1-q/2}R_q^{1/2}M_\Sigma\right)$. Clearly, a faster decay rate of the eigenvalues $\boldsymbol{\lambda}(\Sigma)$ leads to a smaller value of M_Σ, and a faster decay rate of ψ_i's or equivalently a smaller value of R_q, consequently a better rate. Different from the fixed $(x, \mathbf{X})$ considered in Theorem 2, $(x, \mathbf{X})$ are random variables in this section. Thus, R_q lacks an explicit rate, due to randomness of the empirical eigenvalues ψ_i's. The magnitudes of ψ_i's, though can be checked from data, are hard to extract in theory generally, according to random matrix theory. To the best of our knowledge, there is no solution for a general case. To obtain an explicit result, we consider two extreme cases: (1) Σ is (approximately) low rank; (2) $\Sigma = I_p$, the least favorable case.

Proposition 7. *Suppose that x and X_i's are independent from $N(0, \Sigma)$. Assume that $n < p$. Taking $\Gamma = \Gamma_{\text{eg}}$, we have the following conclusions:*

(1) If Σ is of low rank with rank$(\Sigma) = r_\Sigma$, we have $|\hat{\gamma}_x - \gamma_x| = O_p(r_\Sigma\sqrt{n^{-1}\log n})$. An extension to Σ being approximately low rank is presented in Proposition D.1 of Supplementary Material.

(2) For the least favorable case of $\Sigma = I_p$, it holds that $|\hat{\gamma}_x - \gamma_x| = O_p(p^{1/2}(\log n)^{1/4})$.

For the case of Σ having a low rank, $|\hat{\gamma}_x - \gamma_x|$ has the order similar to that of $|\hat{\alpha}_x - \alpha_x|$. But when $\Sigma = I_p$, the error diverges, which is not surprising since $\boldsymbol{\theta}$ is non-sparse in the transformed model in this setting and the rate is determined by the most difficult case. Particularly, the rate for $\Sigma = I_p$ is the combination of the facts that $|\hat{\alpha}_x - \alpha_x| = O_p((\log n)^{1/4})$ and $M_\Sigma = p^{1/2}$. Next we briefly compare LASSO with the proposed method with $\Gamma = \Gamma_{\text{eg}}$. LASSO performs well in prediction when $\boldsymbol{\beta}$ is (approximately) sparse, and is less sensitive to the sparsity of eigenvalues. In contrast, the proposed method with $\Gamma = \Gamma_{\text{eg}}$ has good performance when the eigenvalues of Σ decrease fast, and $\boldsymbol{\beta}$ can be sparse or less sparse. Thus our method with $\Gamma = \Gamma_{\text{eg}}$ and LASSO are complementary to each other.

4.2.2 Error of $\hat{\gamma}_{\tilde{x}_{S_1}}$ in prediction with $\Gamma = \Gamma(\hat{\boldsymbol{\beta}})$

Recall that in Example 2 in Section 2.1, $\gamma_x = \gamma_{\tilde{x}_{S_1}}$ for any $S_1 \supseteq S_0$ with $S_0 = \text{supp}(\boldsymbol{\beta})$, implying that one can make prediction at the point $\tilde{x}_{S_1}$. Trivially, one can take $S_1 = \{1, \cdots, p\}$ such that $\tilde{x}_{S_1} = x$. The subset S_1 takes the sparsity degree of $\boldsymbol{\beta}$ into account. Given an initial estimator $\hat{\boldsymbol{\beta}}$ and a subset S_1 such that $S_1 \supseteq S_0$, by applying our approach with $\Gamma = \Gamma(\hat{\boldsymbol{\beta}})$ that is constructed with x replaced by $\tilde{x}_{S_1}$, we obtain the estimator of $\gamma_{\tilde{x}_{S_1}}$ denoted as $\hat{\gamma}_{\tilde{x}_{S_1}}$.

Denote $d(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = [\max\{\text{var}(X_i^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})), \text{var}(X_{iS_1}^\top(\hat{\boldsymbol{\beta}}_{S_1} - \boldsymbol{\beta}_{S_1}))\}]^{1/2}$, which stands for the prediction error of an initial estimator $\hat{\boldsymbol{\beta}}$. Without loss of generality, we assume $d(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ has magnitude of order no less than $n^{-1/2}$. Let $\Sigma_{S_1S_1} = \text{cov}(X_{iS_1})$ and $M_{S_1} = [tr^2(\Sigma_{S_1S_1})/tr(\Sigma_{S_1S_1}^2)]^{1/2}$ with $M_{S_1}^2$ standing for the effective rank of matrix $\Sigma_{S_1S_1}$. Let $\tilde{H}_{\min}$ be the quantity defined similar to $H_{\min}$ but with the eigenvalues of $n^{-1}\mathbf{X}\mathbf{X}^\top$, satisfying $1 \lesssim \tilde{H}_{\min} \lesssim \sqrt{n}$ (the detailed expression is given in the Supplementary Materials). We have the following conclusions from Theorem 3.

Theorem 4. *Let $\Gamma = \Gamma(\hat{\boldsymbol{\beta}})$. Assume that (i) x and X_i's are i.i.d. variables from $N(0, \Sigma)$*

and $n < p$; (ii) Both S_1 and $\hat{\boldsymbol{\beta}}$ are independent of $(\mathbf{X}, \mathbf{Y})$ satisfying $\operatorname{cov}(X_i^\top \hat{\boldsymbol{\beta}}) \asymp 1$. Then it holds that $|\hat{\gamma}_{\tilde{x}_{S_1}} - \gamma_x| = O_p(\lambda_n M_{S_1} \tilde{H}_{\min} d(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}))$. If we further assume the complementary condition: $d(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}))\tilde{H}_{\min} = O_p(1)$, it holds that $|\hat{\gamma}_{\tilde{x}_{S_1}} - \gamma_x| = O_p(\lambda_n M_{S_1})$. These conclusions are still valid for $S_1 = \{1, \cdots, p\}$.

Theorem 4 shows that the rate depends on $M_{S_1}, \tilde{H}_{\min}$ and $d(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$. The first two depend on the decay rate of eigenvalues. Moreover, it holds that $d(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = O_p(1)$ by the assumptions that both $\operatorname{cov}(X_i^\top \boldsymbol{\beta})$ and $\operatorname{cov}(X_i^\top \hat{\boldsymbol{\beta}})$ are bounded, which imposes restrictions on $\hat{\boldsymbol{\beta}}$. Hence, by Theorem 4, without the complementary condition, we have the error rate $\min\{\lambda_n M_{S_1} \sqrt{n} d(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}), \lambda_n M_{S_1} \tilde{H}_{\min}\}$.

We compare P_{lasso} with the plug-in method using LASSO estimator $\hat{\boldsymbol{\beta}}$ (briefly named LASSO). By the typical rate of LASSO, we have $d(\hat{\boldsymbol{\beta}}_{\text{lasso}}, \boldsymbol{\beta}) = O_p(\min\{1, \sqrt{s_0 \log p/n}\})$. To simplify the comparison, we assume that the LASSO estimator is consistent, that is, $\sqrt{s_0 \log p/n} = o(1)$. Then one can see that the condition $d(\hat{\boldsymbol{\beta}}_{\text{lasso}}, \boldsymbol{\beta})\tilde{H}_{\min} = O_p(1)$ becomes $\tilde{H}_{\min} = O(n^{1/2}(s_0 \log p)^{-1/2})$, which is mild, since it always holds that $\tilde{H}_{\min} \lesssim n^{1/2}$.

Denote by $T_{n,\text{rad}}$ the ratio of the error rate of P_{lasso} over $d(\hat{\boldsymbol{\beta}}_{\text{lasso}}, \boldsymbol{\beta})$ for random test points. Then $T_{n,\text{rad}} = o_p(1)$ would imply that our method is better than LASSO. By Theorem 4, we see that $T_{n,\text{rad}} \asymp \lambda_n M_{S_1} \tilde{H}_{\min}$. To investigate the magnitude of the latter, we consider the following two cases:

- Suppose that $\boldsymbol{\beta}$ is sparse with support set S_0 of cardinality $s_0 = |S_0|$. Moreover, if a good S_1 is available such that $S_1 \supseteq S_0$ and $|S_1| \asymp s_0$, that is, we know sufficiently well on the support set. Then we have $M_{S_1} \asymp s_0^{1/2}$. If $\tilde{H}_{\min} = o_p(n^{1/2}(s_0 \log n)^{-1/2})$ (i.e. $\lambda_n s_0^{1/2} \tilde{H}_{\min} = o_p(1)$) which is mild as argued above, we have $T_{n,\text{rad}} = o_p(1)$; otherwise $T_{n,\text{rad}} = \Omega_p(1)$. Moreover, P_{lasso} has error $|\hat{\gamma}_{\tilde{x}_{S_1}} - \gamma_x| = O_p\left(\sqrt{s_0(\log n)/n}\right)$ under the mild condition $H_{\min} d(\hat{\boldsymbol{\beta}}_{\text{lasso}}, \boldsymbol{\beta}) = O_p(1)$. If one knows the support set S_0 in advance, the OLS estimator using the oracle predictor X_{iS_0}'s has the rate $\sqrt{s_0/n}$, which is similar to the rate of our estimator up to a term $\log n$. However, the performance of our estimator depends on that of S_1 and the decay of eigenvalues.

- If we simply set $S_1 = \{1, \cdots, p\}$, ignoring the sparsity information of $\boldsymbol{\beta}$, then $T_{n,\text{rad}} = o_p(1)$ if and only if $M_{S_1}\tilde{H}_{\min} = o_p(\lambda_n^{-1})$; otherwise $T_{n,\text{rad}} = \Omega_p(1)$. This condition $M_{S_1}\tilde{H}_{\min} = o_p(\lambda_n^{-1})$ holds when eigenvalues decay fast. Moreover, under the settings similar to Corollary 2, we have $\tilde{H}_{\min} = O_p(1)$. For instance, if $rank(\Sigma) = o(\lambda_n^{-1})$, then $M_{S_1}\tilde{H}_{\min} \lesssim rank(\Sigma) = o_p(\lambda_n^{-1})$. However, P_{lasso} is worse than LASSO when Σ is close to I_p and $\boldsymbol{\beta}$ is indeed sparse; specifically, under the complementary condition, P_{lasso} has error rate $\lambda_n M_{S_1} = O_p((n^{-1}p\log n)^{1/2})$, which is worse than that of LASSO. However, taking $S_1 = \{1, \cdots, p\}$ accommodates both sparse and non-sparse $\boldsymbol{\beta}$. The classical OLS estimator for a non-sparse $\boldsymbol{\beta}$ has the order $\sqrt{p/n}$, close to that of P_{lasso}. In practice, we do not know the sparsity degree of $\boldsymbol{\beta}$. The CV approach in Section 2.3 can be used to select between $\hat{\gamma}_{x_{S_1}}$ and $\hat{\gamma}_x$ in an automatic data adaptive manner.

5 Numerical Studies

We use simulation studies in Section 5.1 and real data analysis in Section 5.2 to further illustrate the numerical performance of our method.

5.1 Simulations

We consider the simulation studies with samples generated *i.i.d.* from the linear model (2.1) with $p = 1000$ dimensional vector $X_i \sim N(0, \Sigma)$ and $\epsilon_i \sim N(0, 1)$. Set $\Sigma = (\sigma_{ij})$ with $\sigma_{ij} = 0.5^{|i-j|/\eta}$, where η controls the level of dependence strength among the predictors, with larger values of η implying stronger correlations among predictors. For the convenience of discussion, the plug-in estimator is named by the method used in estimating $\boldsymbol{\beta}$. For

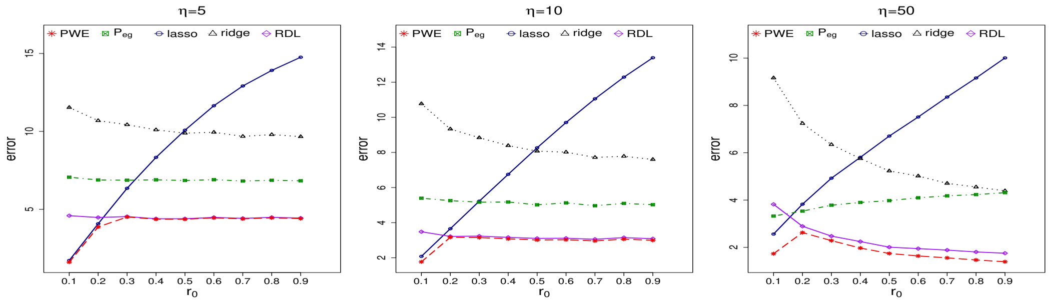

Figure 2: Simulation prediction results of PWE, LASSO, ridge, P_{eg} and ridgeless (RDL).

example, "LASSO", "ridge" and "ridgeless" denote the plug-in estimators $\hat{\boldsymbol{\beta}}^\top x$ with $\hat{\boldsymbol{\beta}}$ being LASSO, ridge and ridgeless estimators, respectively.

Setting 1 (Prediction). We set $\boldsymbol{\beta} = \delta_0(\mathbf{1}_{p_0}^\top, 0, \cdots, 0)^\top \in \mathbb{R}^p$, where $p_0 = r_0 p$, $\mathbf{1}_{p_0}$ is the p_0-dimensional vector of 1, and $\delta_0 = 10/\sqrt{p_0}$ such that $\|\boldsymbol{\beta}\| = 10$. Clearly, $\boldsymbol{\beta}$ is denser for larger values of r_0, and we set $r_0 \in \{0.1, 0.2, \ldots, 0.9\}$. For prediction, we set $\mathcal{M} = \{P_{\text{lasso}}, P_{\text{ridge}}, P_{\text{eg}}, P_{\text{rdl}}\}$ for Algorithm 2 in Section 2.3.

We compare the prediction performance of different methods. Split the data into two parts with the training sample of size $n_{tr} = 200$ and the testing sample of size $n_{te} = 500$ to compute the test error. PWE estimators are obtained by Algorithm 2 with the $\mathcal{M}$ given above and $S_1 = \{1, \cdots, p\}$. For the implementation of Algorithm 1, the bias correction step is adopted. We repeat the procedure 100 times and calculate the average test error for each method. We compare LASSO, ridge, ridgeless, P_{eg} and the PWE estimators. For clarity, simulation results for Setting 1 are summarized as follows:

(1) Comparison of LASSO, ridge, ridgeless with PWE and P_{eg} on prediction. The simulation results are presented in Figure 2. When $\boldsymbol{\beta}$ is sparse such as $r = 0.1$, PWE performs similar to (with small η) and better than (with large η) LASSO, and much better than other methods including ridgeless. When r_0 is large, PWE is similar to or slightly better than the ridgeless estimator, and is much better than other methods. By taking the advantages of different initial estimators, PWE performs well for both sparse and dense

$\boldsymbol{\beta}$. Moreover, P_{eg} is insensitive to r_0, which supports our theoretical findings, and is better than LASSO when r_0 is large. Its advantage over LASSO is more clear when η is large.

(2) Comparison of plug-in estimators with our proposed pointwise counterparts. Due to the limited space, the results are presented in Figure S.2 in Section G.1 of the Supplementary material. It is seen that the performance of P_{lasso} is similar to LASSO for small η and is better than LASSO for large η; in all cases, P_{ridge} is better than ridge. The numerical results match with our theoretical findings in the sense that the sparsity in eigenvalues can be helpful. In addition, P_{rdl} is close to ridgeless, especially when η is small. More comparisons are presented in Setting 4 and Section G.1 in Supplementary Materials.

Setting 2 (Sparse linear transformation). We consider x being sparse vectors in $\gamma_x = \boldsymbol{\beta}^\top x$. Set $\boldsymbol{\beta} = (3, -3, 3, 1, \delta_1 \mathbf{1}_{p_0-4}^\top, \mathbf{0}_{p-p_0}^\top)^\top$, where $\delta_1 = 5/\sqrt{p_0}$. Consider γ_x being one of the four quantities β_1, β_{p_0}, β_p and $\beta_1 - \beta_3$, corresponding to taking x being e_1, e_{p_0}, e_p and $e_1 - e_3$ respectively in γ_x. Note that $\beta_1 - \beta_3 = 0$, indicating that there is no difference between effects of the first and third predictors; $\beta_1 = 3$ and $\beta_{p_0} = \delta_1$ stand for strong and weak signals, respectively. Moreover, $\beta_p = 0$ indicates that the p-th predictor is insignificant. Let $p_0 = 300$ and $\Sigma = (\sigma_{ij})$ with $\sigma_{ij} = 0.5^{|i-j|/150}$, so that the predictors are highly correlated. We set the training data size $n_{tr} = 150$, and compute the average errors of $|\hat{\gamma}_x - \gamma_x|$ over 100 replications. We compare the regularized estimators, P_{eg}, P_{lasso} and P_{ridge}, with the plug-in ones. Estimation comparison results are presented in Table 2.

γ_x	LASSO	A-LASSO	ridge	P_{lasso}	P_{ridge}	P_{eg}	PWE
β_1	0.671	0.681	0.672	0.701	0.405	0.277	0.405
β_{p_0}	0.072	0.072	0.036	0.072	0.072	0.128	0.072
β_p	0	0	0.010	0	0	0	0
$\beta_1 - \beta_3$	0.251	1.062	0.005	0	0	0.022	0

Table 2: The average values of $|\hat{\gamma}_x - \gamma_x|$. The PWE corresponds to the estimator automatically selected from $\{P_{\text{eg}}, P_{\text{lasso}}, P_{\text{ridge}}\}$; A-LASSO stands for the adaptive LASSO.

For a strong signal β_1, it can be inferred from Table 2 that the regularized pointwise estimators P_{ridge} and P_{eg} are better than LASSO, adaptive LASSO and ridge regression. For weak signal β_{p_0}, ridge estimator is better than others. The main reason is that other methods sometimes shrink the estimators to zero, resulting in large biases. For $\beta_p = 0$, except ridge regression, other methods give zero estimates. Finally, for $\beta_1 - \beta_3 = 0$, all the regularized pointwise estimators result in exactly zero estimates, while the plug-in estimators LASSO and adaptive LASSO lead to large biases.

Setting 3 (Comparison on different subset S_1). As pointed out in Sections 2.1 and 2.3, one can consider prediction at the point $\tilde{x}_{S_1}$ instead of x in prediction problems. Under the setup of Setting 1, we take $\Gamma = \Gamma(\hat{\boldsymbol{\beta}}_{\text{lasso}})$ and compare the following four candidates of S_1: (1) S_1 being $S_0 = \text{supp}(\boldsymbol{\beta})$, which is the ideal case; (2) S_1 being $S_{\text{full}} = \{1, \cdots, p\}$, that is, $\gamma_{\tilde{x}_{S_1}} = \gamma_x$; (3) S_1 is obtained by SIS of Fan and Lv (2008), denoted as S_{SIS}; (4) S_1 being the $S_{\text{lasso}} = \text{supp}(\hat{\boldsymbol{\beta}}_{\text{lasso}})$. For each candidate of S_1, we repeat 100 times and report the average values of the prediction error, true positive rate (TP) and the average length (LEN) that are defined as TP$=|S_1 \cap S_0|/|S_0|$ and LEN$=|S_1|/p$ respectively.

Due to the limited space, we present the simulation results in Section G.2 in Supplementary Materials. It is seen that S_0 always leads to the best prediction errors in all cases. When $r_0 = 0.01$ where $\boldsymbol{\beta}$ is very sparse, both S_{SIS} and S_{lasso} have higher values of TP and smaller values of LEN, leading to smaller prediction errors than those of S_{full}. As r_0 increases, the signal of β_j's becomes weak due to the constraint $\|\boldsymbol{\beta}\| = 10$, and the values of TP for S_{SIS} are very small and are the smallest ones among all subsets, which lead to the worst prediction errors. On the other hand, TP and consequently errors of S_{lasso} are much better than those of S_{SIS}, because LASSO takes into account correlations among the predictors when selecting the significant variables, while SIS uses only the marginal correlations. In addition, it is observed that S_{lasso} performs similar to that of S_{full}, which is partially due to the following reason. During the construction of $\Gamma(\hat{\boldsymbol{\beta}}_{\text{lasso}})$ with a given

S_1, we need to compute $\tilde{x}_{S_1}^\top \hat{\boldsymbol{\beta}}_{\text{lasso}}$, which equals $x^\top \hat{\boldsymbol{\beta}}_{\text{lasso}}$ for S_1 being both S_{lasso} and S_{full}.

Setting 4 (*Further comparison for heterogeneous test points*). In Setting 1 where test points are *i.i.d.* copies from the training distribution, P_{lasso} is nearly the same as LASSO when η is small such as $\eta = 5$ (results shown in Figure S.2 in Supplementary Material). We compare them further for the case of x_i's following a distribution different from X_i's, which is known as covariate shift in the literature of transfer learning (Weiss et al., 2016). We generate training data of size 100 as in Setting 1 with $\eta = 5$ and $\boldsymbol{\beta} \propto (\mathbf{1}_{p_0}^\top, 0, \cdots, 0)^\top \in \mathbb{R}^p$ with $\|\boldsymbol{\beta}\| = 5$. The test points x_i's are *i.i.d.* from $N(0, \Sigma_{\text{te}})$. The eigenvectors matrix U_{te} of Σ_{te} is uniformly distributed on the set of all orthogonal matrices in $\mathbb{R}^{p\times p}$. The eigenvalues of Σ_{te}, denoted as $\varrho_{\text{te},1}, \cdots, \varrho_{\text{te},p}$, satisfy that $\varrho_{\text{te},i} = 2(p-i+1)/(p+1)$ such that $tr(\Sigma_{\text{te}}) = p$. We first generate a Σ_{te} and then 200 test points with given Σ_{te}, and repeat this procedure 100 times to compute average prediction errors. Results in Table 3 show that P_{lasso} is much better than LASSO for the case of covariate shift even for small η, possibly due to the flexible pointwise prediction of our proposed method. Besides these examples, additional results demonstrate that P_{rdl} can also substantially improve the ridgeless estimator when covariate shift exists for testing data (Section G.1 of Supplementary Materials).

Table 3: Test errors of LASSO and P_{lasso} with $\eta = 5$ for testing points from $N(0, \Sigma_{\text{te}})$

r_0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
LASSO	3.976	5.878	6.749	7.172	7.461	7.569	7.792	8.108	8.164
P_{lasso}	3.278	3.796	4.077	4.197	4.391	4.421	4.469	4.614	4.737

5.2 Real data analysis

We apply our method to a dataset from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (https://adni.loni.usc.edu/). Alzheimer's Disease (AD) is a form of dementia characterized by progressive cognitive and memory deficits. The Mini Mental State Examination (MMSE) is a very useful score in practice for the diagnosis of AD. Generally, any

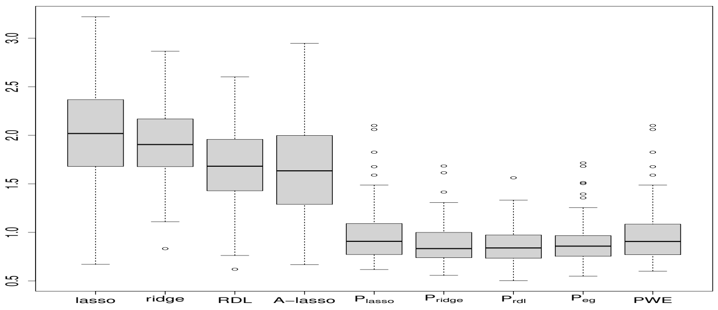

Figure 3: Comparison of the test errors of different methods for the AD data analysis. Here $P_{\text{lasso}}, P_{\text{ridge}}, P_{\text{rdl}}$, and P_{eg} represent our regularized estimators. The PWE is automatically selected from $P_{\text{lasso}}, P_{\text{ridge}}, P_{\text{rdl}}$, and P_{eg}.

score greater than or equal to 27 points (out of 30) indicates a normal cognition. Below this, MMSE score can indicate severe ($\leq$ 9 points), moderate (10-18 points) or mild (19-24 points) cognitive impairment (Mungas, 1991). Currently, structural magnetic resonance imaging (MRI) is one of the most popular and powerful techniques for the diagnosis of AD. One can use MRI data to predict the MMSE score and identify the important diagnostic and prognostic biomarkers. The dataset we used contains the MRI data and MMSE scores of 51 AD patients and 52 normal controls. After the image preprocessing steps for the MRI data, we obtain the subject-labeled image based on a template with 93 manually labeled regions of interest (ROI) (Zhang and Shen, 2012). For each of the 93 ROI in the labeled MRI, the volume of gray matter tissue is used as a feature. Therefore, the final dataset has 103 subjects. For each subject, there are one MMSE score and 93 MRI features. We treat the MMSE score as the response variable and MRI features as predictors.

We split the data at random with 80% as the training set, denoted as $\mathcal{S}_{tr}$, and 20% as the testing set, denoted as $\mathcal{S}_{te}$, then compute the average test error $\sum_{Y_i \in \mathcal{S}_{te}} |\hat{Y}_i - Y_i|/|\mathcal{S}_{te}|$. We repeat the procedure for 100 times and report the average test errors. The box plots of different methods are presented in Figure 3. It shows that pointwise estimators are much better than plug-in estimators of LASSO, adaptive LASSO and ridge, respectively.

6 Discussion

In this paper, we estimate the linear transformation $\boldsymbol{\beta}^\top x$ of parameters $\boldsymbol{\beta}$ in high dimensional linear models. We propose a pointwise estimator, which works well when $\boldsymbol{\beta}$ is sparse or non-sparse, and predictors are highly or weakly correlated. The theoretical analysis reveals the significant difference between estimating a linear transformation of $\boldsymbol{\beta}^\top x$ and that of $\boldsymbol{\beta}$. When $\boldsymbol{\beta}$ is non-sparse or predictors are highly correlated, estimating $\boldsymbol{\beta}$ is difficult, but we can still get good estimate of $\boldsymbol{\beta}^\top x$ using our proposed pointwise estimators.

References

Azriel, D. and A. Schwartzman (2020). Estimation of linear projections of non-sparse coefficients in high-dimensional regression. *Electronic Journal of Statistics 14*(1), 174–206.

Bartlett, P. L., P. M. Long, G. Lugosi, and A. Tsigler (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences 117*, 30063 – 30070.

Belloni, A. and V. Chernozhukov (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli 19*(2), 521–547.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of statistics 37*(4), 1705–1732.

Bühlmann, P. and S. Van De Geer (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Cai, T. T. and Z. Guo (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of statistics 45*(2), 615–646.

Dalalyan, A. S., M. Hebiri, and J. Lederer (2017). On the prediction performance of the lasso. *Bernoulli 23*(1), 552–581.

Dobriban, E. and S. Wager (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics 46*(1), 247–279.

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

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

Hastie, T., A. Montanari, S. Rosset, and R. J. Tibshirani (2022). Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics 50*(2), 949–986.

Hebiri, M. and J. Lederer (2013). How correlations influence lasso prediction. *IEEE Transactions on Information Theory 59*(3), 1846–1854.

Javanmard, A. and A. Montanari (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research 15*(1), 2869–2909.

Lu, S., Y. Liu, L. Yin, and K. Zhang (2017). Confidence intervals and regions for the lasso by using stochastic variational inequality techniques in optimization. *Journal of the Royal Statistical Society: Series B 79*(2), 589–611.

Mungas, D. (1991). In-office mental status testing: a practical guide. *Geriatrics 46*(7), 54–67.

Negahban, S. N., P. Ravikumar, M. J. Wainwright, and B. Yu (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science 27*(4), 538–557.

Raskutti, G., M. J. Wainwright, and B. Yu (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q -balls. *IEEE Transactions on Information Theory 57*(10), 6976–6994.

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

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics 42*(3), 1166–1202.

Weiss, K., T. M. Khoshgoftaar, and D. Wang (2016). A survey of transfer learning. *Journal of Big data 3*(1), 1–40.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics 38*(2), 894–942.

Zhang, C. H. and S. S. Zhang (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society Series B 76*(1), 217–242.

Zhang, D. and D. Shen (2012). Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in alzheimer's disease. *Neuroimage 59*(2), 895–907.

Zhang, Y., J. Duchi, and M. Wainwright (2015). Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *The Journal of Machine Learning Research 16*(1), 3299–3340.

Zhang, Y., M. J. Wainwright, and M. I. Jordan (2017). Optimal prediction for sparse linear models? lower bounds for coordinate-separable m-estimators. *Electronic Journal of Statistics 11*(1), 752–799.

Zhu, Y. and J. Bradic (2018). Linear hypothesis testing in dense high-dimensional linear models. *Journal of the American Statistical Association 113*(524), 1583–1600.

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