Statistica Sinica Preprint No: SS-2025-0303	
Title	Linear Shrinkage Convexification of Penalized Linear Regression with Missing Data
Manuscript ID	SS-2025-0303
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0303
Complete List of Authors	Seongoh Park, Seongjin Lee, Nguyen Thi Hai Yen, Nguyen Phuoc Long and Johan Lim
Corresponding Authors	Johan Lim
E-mails	johanlim@snu.ac.kr
Notice: Accepted author version.	

Statistica Sinica

Linear Shrinkage Convexification of Penalized Linear Regression With Missing Data

Seongoh Park[1], Seong Jin Lee[2], Nguyen Thi Hai Yen[3],

Nguyen Phuoc Long[4], Johan Lim[5,*]

[1]*Sungshin Women's University,* [2]*University of North Carolina at Chapel Hill,*

[3]*Inje University College of Medicine,* [4]*Chang Gung University,* [5]*Seoul National University*

Abstract:

One of the common challenges faced by researchers in recent data analysis is missing values. In the context of penalized linear regression, which has been extensively explored over several decades, missing values introduce bias and yield a non-positive definite covariance matrix of the covariates, rendering the least square loss function non-convex. In this paper, we propose a novel procedure called the linear shrinkage positive definite (LPD) modification to address this issue. The LPD modification aims to modify the covariance matrix of the covariates in order to ensure consistency and positive definiteness. Employing the new covariance estimator, we are able to transform the penalized regression problem into a convex one, thereby facilitating the identification of sparse solutions. Notably, the LPD modification is computationally efficient and can be expressed analytically. In the presence of missing values, we establish the selection consistency and prove the convergence rate of the ℓ_1-penalized regression estimator with LPD, showing an ℓ_2-error convergence rate of square-root of $\log p$ over n by a factor of $(s_0)^{3/2}$ (s_0: the number of non-zero coefficients). To further evaluate the effectiveness of our approach, we analyze real data from the Genomics of Drug Sensitivity in Cancer (GDSC) dataset. This dataset provides incomplete measurements of drug sensitivities of cell lines and their protein expressions. We conduct a series of penalized linear regression models with each sensitivity value serving as a response variable and protein expressions as explanatory variables.

*To whom all correspondence should be addressed.

Key words and phrases: general missing dependency, lasso, positive definiteness.

1. Introduction

Regularized or penalized linear regression has been largely explored for decades, motivated from a variety of modern applied fields (Lee et al., 2003; Ghosh and Chinnaiyan, 2005; Daye et al., 2012; Han and Tsay, 2020) where the sample size is much smaller than the number of variables to be analyzed. Among different regularizations in linear regression such as ridge (Hoerl and Kennard, 1970), lasso (Tibshirani, 1996; Zou, 2006), Dantzig selector (Candes and Tao, 2007), elastic net (Zou and Hastie, 2005), SCAD (Fan and Li, 2001), the lasso regression has gained its popularity because its statistical properties (Zhao and Yu, 2006; Lee et al., 2015; Zou, 2006; van de Geer and Bühlmann, 2009; Fu and Knight, 2000) and computational aspects (Efron et al., 2004; Osborne et al., 2000; Friedman et al., 2007) are well established.

Though the technology for data collection has exceptionally advanced in recent years, one common issue that researchers face in data analyses is missing values. Our motivating example is drug response data (`https://www.cancerrxgene.org/`, Release v8.4, July 2022) and the pan-cancer proteomic profile of 8,498 proteins from 949 human cancer cell lines (28 tissue types, more than 40 cancer types) (Gonçalves et al., 2022). This study was to measure the sensitivities (IC50/AUC) of cells to different drugs and aimed to find the association between drug responses and protein levels. Missing data are widely seen in mass spectrometry (MS)-based proteomics (Webb-Robertson et al., 2015) or metabolomics (Wei et al., 2018). Causes for missing values could be biological or technical (e.g., stochastic fluctuations during data acquisition) and of random or not at-random (Karpievitch et al.,

2012). Unless treated appropriately, incomplete data often lead to biased results and hamper study reproducibility (Dabke et al., 2021). For instance, for the lasso regression Sørensen et al. (2015) showed that a naive approach using the incomplete data without correction does not satisfy estimation consistency (see Proposition 1 therein).

Many researchers have come up with different solutions to address this issue under linear regression models. First, the expectation-maximization (EM) algorithm is developed by Stdler and Bhlmann (2010) where they aimed to find the sparse inverse covariance matrix and used it in the sparse linear regression. However, the EM algorithm is model-specific and known to converge slowly. Alternatively, variable selection can be combined with multiple imputation that is commonly used in practice. For example, one can perform majority votes based on selection results from multiply imputed datasets (Heymans et al., 2007; Wood et al., 2008; Lachenbruch, 2011; Long and Johnson, 2015). To avoid the ad-hoc rules for combining different sets of selected variables, Wan et al. (2015) and Li et al. (2023) considered stacking imputed datasets and selected the same variables across all datasets, which is termed as a stacked method in Du et al. (2022). In Chen and Caramanis (2013), they proposed the group-wise selection approach to consistently choose variables across imputed datasets, which is named a grouped method in Du et al. (2022). These methods exhibited satisfactory performance in simulated and real data analyses; however, theoretical evidences are elusive.

To fill this gap, researchers have paid attention on de-biasing approaches. These are based on the observation that a loss function, for example, mean squared error, is biased if data are not completely observed. Thus, related work adjusted it by adding or multiplying de-biasing constants to the covariance part or Gram matrix (e.g. see (2.5)) and solved the corrected optimization problem with different penalization methods; for example, Liang and

Li (2009) used the SCAD penalty, and Loh and Wainwright (2012) adopted the lasso penalty. Following Loh and Wainwright (2012) where estimation consistency is proved, Sørensen et al. (2015) additionally showed sign consistency under the irrepresentable condition adapted to their contexts. This line of work, however, has a computational issue that the modified loss function is no longer convex. It was sidestepped in Rosenbaum and Tsybakov (2010) and Wang et al. (2019) by using Dantzig selector that is always defined as a linear programming regardless of the modification.

A more fundamental remedy for the non-convexity is to modify the corrected covariance factor $\widehat{\mathbf{\Sigma}}$ to be positive definite (PD). To this end, Datta and Zou (2017) found the closest PD matrix to $\widehat{\mathbf{\Sigma}}$ using the element-wise maximum norm:

$$\tilde{\mathbf{\Sigma}}^{CoCo} = \arg\min_{\mathbf{\Sigma} \succ 0} \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max}. \tag{1.1}$$

Using it, they solved the ℓ_1-penalized regression problem, which is named CoColasso, and proved estimation and selection consistency under regular conditions including the irrepresentable condition. This area of research has been recently studied further. Though handling the measurement error not missing data, Zheng et al. (2018) and Zhang et al. (2022) proposed to use different penalty functions, a combination of ℓ_1- and concave penalty, and ℓ_0-penalty, respectively, to ensure better theoretical properties of estimators (i.e. faster oracle inequality). Escribe et al. (2021) considered partially corrupted data where some of explanatory variables are corrupted under some measurement error model and the others are not. Thus, they only solved (1.1) for a smaller dimension at which the measurement errors are found. On the other hand, in solving (1.1), Takada et al. (2019) suggested to downweight components at which samples are highly missing. To do so, they used a weighted version of Frobenius norm.

However, solving (1.1) is computationally demanding in general because it does not have a closed form solution. More specifically, the eigen-decomposition of a p-dimensional symmetric matrix and projection of a p^2-dimensional vector to ℓ_1-ball are repeated until convergence (Datta and Zou, 2017; Han et al., 2014). Takada et al. (2019) used the (weighted) Frobenius norm to find the closest PD matrix in which the eigen-decomposition is also repeated. Because of this, the existing methods mentioned above may not be practically useful. The heavy workload can greatly impede further inference procedures using regularized estimators such as bolasso (bootstrapped enhanced lasso, Bach (2008)) and a modified residual bootstrapped lasso, which are based on resampling procedures (Chatterjee and Lahiri (2011, 2013) or stability selection (Meinshausen and Bhlmann, 2010)). Moreover, there is a need for solving the penalized regression recursively; e.g. online learning procedure (Duchi and Singer, 2009; Langford et al., 2008; Xiao, 2009).

In this paper, we propose the linear shrinkage positive definite (LPD) modification of the covariance matrix for the high-dimensional regression problem with incomplete data. The key idea is to reduce the class of PD matrices over which the minimization (1.1) is taken. We consider the linear shrinkage class defined in (2.8). In other words, we shrink the non-PD $\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}$ (corrected estimator defined in (2.5)) to $\mu\mathbf{I}$ as $\alpha\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}} + (1-\alpha)\mu\mathbf{I}$ for some α and μ. The proposed way is easy and straightforward due to its simple form, and above all, computationally fast since the optimal α and μ have explicit forms (see (2.10) and Proposition 2). Based on the new covariance estimators, we convexify the penalized regression problem and thus can easily find the sparse solution $\widehat{\boldsymbol{\beta}}^{\mathrm{LPD}}$ to (2.7). Furthermore, under the irrepresentable condition, we establish the selection consistency and prove the rate of convergence by $O_p\left(\sqrt{\log p/n}\right)$ in ℓ_2-error, which is comparable to what was previously

achieved by CoColasso (Datta and Zou, 2017). One of the key tools to prove the results is the non-asymptotic inequality of the IPW estimator (Theorem S1 in Supplementary Materials S1), which can be of independent interest. Our numerical study also reveals the proposed one performs comparatively in the finite sample scenarios. We also analyze real data from Genomics of Drug Sensitivity in Cancer (GDSC) where sensitivity to different drugs and protein expressions was measured but incompletely. We separately run a list of penalized linear regression models with each of sensitivity values as a response variable and protein expressions as explanatory variables, which would have not been feasible if our estimation procedure were not scalable like CoColasso.

The remainder of the paper is organized as follows. In Section 2, we define different classes of linear shrinkage estimators from different matrix norms. Then, we describe how to use the modified Gram matrix in the lasso regression and verify theoretical properties of the resulting lasso estimator under some conditions. In Section 3, we examine the finite sample performance of the proposed method compared to existing methods through simulated data. In Section 4, the proposed regularized regression is applied to incomplete data from Genomics of Drug Sensitivity in Cancer (GDSC) to identify the most predictive proteins for two example drugs. In Section 5, we conclude this paper with a discussion of limitations and potential extensions.

2. Convexification of Lasso using LPD

2.1 Problem formulation

We assume a linear relationship between explanatory variables $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ and a response variable y_i, which is represented by regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$:

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, n, \tag{2.2}$$

where ϵ_i is an error term independent of $\boldsymbol{x}_i$, and samples are independent across $i = 1, \ldots, n$. For ease of exposition, we assume all the variables are centered; $\mathbb{E}x_{ij} = \mathbb{E}\epsilon_i = 0$ and thus $\mathbb{E}y_i = 0$. Due to the missing structure, we can only observe $\tilde{y}_i, \tilde{\boldsymbol{x}}_i = (\tilde{x}_{i1}, \ldots, \tilde{x}_{ip})^\top$ where

$$\tilde{y}_i = \begin{cases} y_i, & \text{if } y_i \text{ is observed,} \\ 0, & \text{otherwise,} \end{cases} \qquad \tilde{x}_{ij} = \begin{cases} x_{ij}, & \text{if } x_{ij} \text{ is observed,} \\ 0, & \text{otherwise.} \end{cases} \tag{2.3}$$

Adopting matrix notations, we write $\tilde{\boldsymbol{y}} = (\tilde{y}_1, \ldots, \tilde{y}_n)^\top$ and $\tilde{\boldsymbol{X}} = [\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_n]^\top$. The penalized regression problem of our interest would be defined by minimizing the residual sum of squares, $\min_{\boldsymbol{\beta}} \frac{1}{2n}\|\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}\|_2^2 + J_\lambda(\boldsymbol{\beta})$ for some penalty function J_λ indexed by a tuning parameter $\lambda > 0$. The problem can be depicted with covariance terms, $\boldsymbol{S} = \tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}/n$ and $\boldsymbol{r} = \tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{y}}/n$, i.e.

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{S}\boldsymbol{\beta} - \boldsymbol{r}^\top \boldsymbol{\beta} + J_\lambda(\boldsymbol{\beta}) \equiv g(\boldsymbol{\beta}; \boldsymbol{S}, \boldsymbol{r}, J_\lambda). \tag{2.4}$$

However, bias caused by missing values in $\boldsymbol{S}$ and $\boldsymbol{r}$ renders the optimal solution of the above inconsistent. A straightforward remedy is to adjust the bias through an inverse probability weighting (IPW) technique and to use the corrected estimators: i.e. $\boldsymbol{S} \leftarrow \widehat{\boldsymbol{\Sigma}}^{\mathrm{IPW}}, \boldsymbol{r} \leftarrow \hat{\boldsymbol{\rho}}^{\mathrm{IPW}}$. The IPW estimators are defined by correcting every component with an observation

probability:

$$\widehat{\boldsymbol{\Sigma}}^{\mathrm{IPW}} = \boldsymbol{S} * \left[\frac{1}{\pi_{jk}^{xx}}, 1 \le j,k \le p\right], \quad \hat{\boldsymbol{\rho}}^{\mathrm{IPW}} = \boldsymbol{r} * \left[\frac{1}{\pi_{j}^{xy}}, 1 \le j \le p\right], \tag{2.5}$$

where $*$ is the element-wise product between two matrices (or vectors) of the same size. π_{jk}^{xx} is a probability that the (j,k)-th explanatory variables are observed, and π_{j}^{xy} that the j-th explanatory variable and response variable are observed. They are precisely defined in Assumption 2. The idea of replacing the sample covariances by the IPW estimators has been used in covariance/precision matrix estimation (Park et al., 2023; Lounici, 2014; Park and Lim, 2019; Park et al., 2021; Pavez and Ortega, 2021; Cai and Zhang, 2016). However, $\widehat{\boldsymbol{\Sigma}}^{\mathrm{IPW}}$ is not PD in general, and thus $g(\boldsymbol{\beta}; \widehat{\boldsymbol{\Sigma}}^{\mathrm{IPW}}, \hat{\boldsymbol{\rho}}^{\mathrm{IPW}}, J_\lambda)$ in (2.4) is not convex, even if J_λ is convex (e.g. lasso penalty). Thus, we use a PD alternative based on the linear shrinkage method (Ledoit and Wolf, 2004; Choi et al., 2019), which finds a PD matrix closest to the non-PD in the linear shrinkage class. It solves

$$\Phi_{\mu,\alpha}(\widehat{\boldsymbol{\Sigma}}^{\mathrm{IPW}}) \in \underset{\Phi_{\mu,\alpha} \in \mathcal{C}_\epsilon(\widehat{\boldsymbol{\Sigma}}^{\mathrm{IPW}})}{\mathrm{Arg\,min}} \left\| \widehat{\boldsymbol{\Sigma}}^{\mathrm{IPW}} - \Phi_{\mu,\alpha} \right\|, \tag{2.6}$$

for some matrix norm $\|\cdot\|$, where $\mathcal{C}_\epsilon$ is a class of the linear shrinkage matrices defined in (2.9). Hereafter, we name the PD modification using the linear shrinkage method as LPD and denote the solution $\Phi_{\mu,\alpha}(\widehat{\boldsymbol{\Sigma}}^{\mathrm{IPW}})$ by $\widehat{\boldsymbol{\Sigma}}^{\mathrm{LPD}}$ for notational simplicity. In the following sections, we give a detailed account of explicit forms of LPDs in different matrix norms (Section 2.2). In the next section (Section 2.3), we study theoretical properties of the solution of the lasso regression:

$$\min_{\boldsymbol{\beta}} \ \frac{1}{2}\boldsymbol{\beta}^\top \widehat{\boldsymbol{\Sigma}}^{\mathrm{LPD}} \boldsymbol{\beta} - \boldsymbol{\beta}^\top \hat{\boldsymbol{\rho}}^{\mathrm{IPW}} + \lambda \|\boldsymbol{\beta}\|_1, \tag{2.7}$$

where $\widehat{\boldsymbol{\Sigma}}^{\mathrm{LPD}}$ is applied as the Gram matrix.

We end this section by introducing the results of Lee et al. (2015) where the authors study

a generalized framework for the regularized M-estimators that includes our problem (2.7). To prove the rate of convergence in terms of ℓ_2-error and consistent recovery of the support, they assumed three conditions (i) restricted strong convexity (RSC), (ii) irrepresentability condition (IR), and (iii) bounded gradient condition (BG). We refer to Supplementary Materials S2.1 or the original reference for more details about the formulation. In our context, the IR and BG conditions are simplified to the condition (C1) and (C2) of Proposition 2, while the RSC condition is reduced to (C3) of it due to the linear shrinkage structure.

To describe the results, we introduce notations. Consider the model space $M_{\mathcal{A}} = \{\boldsymbol{\beta} \in \mathbb{R}^p : \beta_j = 0, j \in \mathcal{A}^c\}$ where $\mathcal{A} \subset [p]$ is the support of true parameter $\boldsymbol{\beta}^*$. We divide a square matrix using the support $\mathcal{A}$ and denote by $\boldsymbol{A}_{\mathcal{A}\mathcal{A}}, \boldsymbol{A}_{\mathcal{A}\mathcal{A}^c}, \boldsymbol{A}_{\mathcal{A}^c\mathcal{A}}, \boldsymbol{A}_{\mathcal{A}^c\mathcal{A}^c}$, each of which restricts rows and columns of $\boldsymbol{A}$ on corresponding index sets. We denote by $\lambda_{\min}(\boldsymbol{A})$ or $\lambda_{\max}(\boldsymbol{A})$ the smallest or largest eigenvalue of $\boldsymbol{A}$, respectively. Then, we can easily derive the following based on the results in Lee et al. (2015). Remark that the norm in (C1) is the matrix ℓ_∞-norm (i.e. maximum of column-wise sum) and the one in (C2) is the element-wise maximum norm of a vector.

Proposition 1. Assume $\lambda_{\min}(\widehat{\boldsymbol{\Sigma}}^{\mathrm{IPW}}) < 0$. For $\epsilon > 0$ such that $\epsilon < \lambda_{\min}(\boldsymbol{\Sigma})$, define by $\widehat{\boldsymbol{\Sigma}}^{\mathrm{LPD}}$ the LPD of $\widehat{\boldsymbol{\Sigma}}^{\mathrm{IPW}}$ over the class $\mathcal{C}_\epsilon(\widehat{\boldsymbol{\Sigma}}^{\mathrm{IPW}})$. Suppose there exists constants $\tilde{\tau} \in (0, 1)$ and $\lambda > 0$ such that:

$$\text{(C1)} \quad \left\| \widehat{\boldsymbol{\Sigma}}^{\mathrm{LPD}}_{\mathcal{A}^c\mathcal{A}} (\widehat{\boldsymbol{\Sigma}}^{\mathrm{LPD}}_{\mathcal{A}\mathcal{A}})^{-1} \right\|_\infty \leq 1 - \tilde{\tau},$$

$$\text{(C2)} \quad \frac{4(2-\tilde{\tau})}{\tilde{\tau}} \left\| \widehat{\boldsymbol{\Sigma}}^{\mathrm{LPD}} \boldsymbol{\beta}^* - \hat{\boldsymbol{\rho}}^{\mathrm{IPW}} \right\|_\infty < \lambda,$$

$$\text{(C3)} \quad \min_{\boldsymbol{t}:\boldsymbol{t}\neq 0, \boldsymbol{t}_{\mathcal{A}^c}=0} \boldsymbol{t}^\top \widehat{\boldsymbol{\Sigma}}^{\mathrm{LPD}} \boldsymbol{t} / \boldsymbol{t}^\top \boldsymbol{t} \geq \min\{0.5\lambda_{\min}(\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}), \mu\},$$

Then, the followings hold:

(R1) The minimizer $\widehat{\boldsymbol{\beta}}^{\mathrm{LPD}}$ of (2.7) is unique,

(R2) $\|\widehat{\boldsymbol{\beta}}^{\mathrm{LPD}} - \boldsymbol{\beta}^*\|_2 \leq \dfrac{4}{\min\{\lambda_{\min}(\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}), \mu\}} \left(1 + \dfrac{\tilde{\tau}}{4}\right) \sqrt{|\mathcal{A}|}\lambda,$

(R3) $\hat{\beta}_j^{\mathrm{LPD}} = 0, \quad j \in \mathcal{A}^c.$

The proof of Proposition 1 is postponed to Supplementary Materials S2.1, which is offered solely for completeness. We do not assert any contribution to it.

2.2 Explicit forms of LPD

In the estimation of high dimensional covariance matrix (Bickel and Levina, 2008b,a; Rothman, 2012), structural assumptions on true covariance matrix are often made, and many regularized estimators are proposed accordingly. However, the regularization typically does not impose PDness, which makes the resulting estimate not PD in general. Several efforts are made to find an estimator that attains both sparsity and PDness (Bien and Tibshirani, 2011; Lam and Fan, 2009; Liu et al., 2014; Rothman, 2012; Xue et al., 2012; Choi et al., 2019). Among them, the fixed support positive definite modification (FSPD) by Choi et al. (2019) is initially designed to make a covariance matrix estimator PD while preserving its support as its name indicates. However, FSPD is still tempting even for cases where we do not have structural assumptions on covariance matrices but need PDness. Since it is computationally easy and is applicable to any non-PD matrix, we adopt this idea for estimating the PD gram matrix under the missing data structure.

Let $\boldsymbol{A}$ be a real symmetric matrix to be modified PD. For a given $\epsilon > 0$, we define the class of LPD by

$$\mathcal{C}_\epsilon(\boldsymbol{A}) = \{\alpha\boldsymbol{A} + (1-\alpha)\mu\mathbf{I} : \alpha \in (0,1), \mu \in \mathbb{R}, \alpha\lambda_{\min}(\boldsymbol{A}) + (1-\alpha)\mu \geq \epsilon\}. \tag{2.8}$$

Following Choi et al. (2019) and Cho et al. (2021), we minimize a distance induced by any matrix norm $||\cdot||$:

$$\min_{\Phi_{\mu,\alpha}\in\mathcal{C}_\epsilon(\boldsymbol{A})} \|\boldsymbol{A}-\Phi_{\mu,\alpha}\|. \tag{2.9}$$

Note that the minimization is taken over (μ,α), and the distance in (2.6) is indeed rewritten as $\|\alpha\boldsymbol{A}+(1-\alpha)\mu\mathbf{I}-\boldsymbol{A}\| = (1-\alpha)\|\mu\mathbf{I}-\boldsymbol{A}\|$. In the meantime, if $\lambda_{\min}(\boldsymbol{A}) < \epsilon \le \mu$, the constraint can be expressed as

$$\alpha\lambda_{\min}(\boldsymbol{A})+(1-\alpha)\mu \ge \epsilon \iff \alpha \le \frac{\mu-\epsilon}{\mu-\lambda_{\min}(\boldsymbol{A})}.$$

We thus know that the optimal solution α^* for fixed $\mu \ge \epsilon$ is

$$\alpha^* = \alpha^*(\mu) = \frac{\mu-\epsilon}{\mu-\lambda_{\min}(\boldsymbol{A})}. \tag{2.10}$$

regardless of the type of the norm. On the other hand, the solution to μ depends on the distance we use. The following proposition summarizes the results. We define matrix norms as $||\boldsymbol{A}||_2 = \sqrt{\lambda_{\max}(\boldsymbol{A}^\top\boldsymbol{A})}$, $||\boldsymbol{A}||_F = \sqrt{\mathrm{tr}(\boldsymbol{A}^\top\boldsymbol{A})/d_2}$, $||\boldsymbol{A}||_\infty = \max_{i\in[d_1]}\sum_{j=1}^{d_2}|a_{ij}|$, $||\boldsymbol{A}||_{\max} = \max_{i\in[d_1],j\in[d_2]}|a_{ij}|$ for any real matrix $\boldsymbol{A}\in\mathbb{R}^{d_1\times d_2}$.

Proposition 2. For a given symmetric matrix $\boldsymbol{A} = (a_{ij})_{1\le i,j\le p}$ with positive diagonals, assume $\lambda_{\min}(\boldsymbol{A}) < 0 < \epsilon \le \mu$. The linear shrinkage Φ_{μ,α^*} of $\boldsymbol{A}$ achieves the minimum at different values of μ according to different matrix norms.

1. (Spectral norm, Lemma 2 of Choi et al. (2019))

$$\|\boldsymbol{A}-\Phi_{\mu,\alpha^*}\|_2 = \epsilon-\lambda_{\min}(\boldsymbol{A})$$

   for any $\mu \ge \max\{\epsilon, (\lambda_{\max}(\boldsymbol{A})+\lambda_{\min}(\boldsymbol{A}))/2\}$.

2. ((Scaled) Frobenius norm, Lemma 3 of Choi et al. (2019))

$$\left\|\boldsymbol{A}-\Phi_{\mu_F^*,\alpha^*}\right\|_F = (\epsilon-\lambda_{\min}(\boldsymbol{A}))\sqrt{\mu_F^*}$$

where $\mu_F^* = \sum_{j=1}^p (\lambda_j(\boldsymbol{A}) - \bar{\lambda})^2 / \sum_{j=1}^p (\lambda_j(\boldsymbol{A}) - \lambda_{\min}(\boldsymbol{A}))^2$ and $\bar{\lambda}$ is an average of the eigenvalues of $\boldsymbol{A}$, $\lambda_1(\boldsymbol{A}), \ldots, \lambda_p(\boldsymbol{A})$.

3. (ℓ_∞-norm, Lemma 3 of Cho et al. (2021))

$$\|\boldsymbol{A} - \Phi_{\mu,\alpha^*}\|_\infty$$
$$= \begin{cases} \searrow \epsilon - \lambda_{\min}(\boldsymbol{A}) \text{ as } \mu \to \infty, & \text{if } \lambda_{\min}(\boldsymbol{A}) + M_2 > 0, \\ \epsilon - \lambda_{\min}(\boldsymbol{A}), & \text{for any } \mu \geq (M_1 - M_2)/2, \\ & \quad \text{if } \lambda_{\min}(\boldsymbol{A}) + M_2 = 0, \\ (\epsilon - \lambda_{\min}(\boldsymbol{A})) \dfrac{(M_1 + M_2)/2}{(M_1 - M_2)/2 - \lambda_{\min}(\boldsymbol{A})}, & \text{at } \mu = (M_1 - M_2)/2, \\ & \quad \text{if } \lambda_{\min}(\boldsymbol{A}) + M_2 < 0, \end{cases}$$

where $M_1 = \max_j \left(a_{jj} + \sum_{i:i\neq j} |a_{ij}|\right)$ and $M_2 = \max_j \left(-a_{jj} + \sum_{i:i\neq j} |a_{ij}|\right)$. Note that if $\lambda_{\min}(\boldsymbol{A}) + M_2 > 0$, there is no solution.

4. (Element-wise maximum norm)

$$\|\boldsymbol{A} - \Phi_{\mu,\alpha^*}\|_{\max}$$
$$= \begin{cases} \dfrac{(\epsilon - \lambda_{\min}(\boldsymbol{A}))(a_{d,\max} - a_{d,\min})/2}{(a_{d,\max} + a_{d,\min})/2 - \lambda_{\min}(\boldsymbol{A})}, & \text{at } \mu = (a_{d,\max} + a_{d,\min})/2, \\ & \quad \text{if } (a_{d,\max} - a_{d,\min})/2 > a_{\text{off},\max}, \\ \dfrac{(\epsilon - \lambda_{\min}(\boldsymbol{A}))a_{\text{off},\max}}{a_{d,\min} + a_{\text{off},\max} - \lambda_{\min}(\boldsymbol{A})}, & \text{at } \mu = a_{d,\min} + a_{\text{off},\max}, \\ & \quad \text{if } (a_{d,\max} - a_{d,\min})/2 \leq a_{\text{off},\max}. \end{cases}$$

where $a_{d,\max} = \max_j a_{jj}$, $a_{d,\min} = \min_j a_{jj}$, and $a_{\text{off},\max} = \max_{i\neq j} |a_{ij}|$.

We only provide a proof of the last case of Proposition 2, which is in Supplementary Materials S2.2, and for the others we refer readers to the original references.

Finally, we provide a guideline to choose tuning parameters of LPD, i.e. ϵ and μ. Throughout this paper, we set $\epsilon = 10^{-4}$, and the choice has exhibited satisfactory performance. The choice of μ depends on the type of matrix norms in LPD and matrix $A = \widehat{\mathbf{\Sigma}}^{\text{IPW}}$. The result of the Frobenius norm suggests the single best value for μ. In contrast, when the spectral norm is used, any choice of μ beyond $\mu_{lb} := \max\{\epsilon, (\lambda_{\max}(\widehat{\mathbf{\Sigma}}^{\text{IPW}}) + \lambda_{\min}(\widehat{\mathbf{\Sigma}}^{\text{IPW}}))/2\}$ is sufficient for the optimality of shrinkage (see Proposition 2). In this case, we recommend use the value at the boundary because it minimizes the effect of shrinkage. However, the choice is not sensitive in practice, which is verified in our simulation study where different values of μ are compared.

2.3 Main results for consistency

In this section, we check the two conditions in Proposition 1, and compute the convergence rate of $\widehat{\boldsymbol{\beta}}^{\text{LPD}}$ in ℓ_2-norm. Prior to it, we state the assumptions and data structure more precisely.

We introduce binary random variables that indicate whether each entry of data is observed or not: $\delta_i^y = \mathrm{I}(y_i \text{ is observed})$, $\delta_{ij}^x = \mathrm{I}(x_{ij} \text{ is observed})$, $i = 1, \ldots, n$, $j = 1, \ldots, p$. Then, we can concisely express the observed data by the product of the indicator variable and the data, i.e. $\tilde{y}_i = \delta_i^y y_i, \tilde{x}_{ij} = \delta_{ij}^x x_{ij}$, which is equivalent to (2.3).

We define the sub-Gaussian (or ψ_2-) norm of a random variable X in $\mathbb{R}$ by $||X||_{\psi_2} = \sup_{p \geq 1}(\mathbb{E}|X|^p)^{1/p}/\sqrt{p}$, and X is called sub-Gaussian if its ψ_2-norm is bounded. Under the regression setting (2.2), we assume the following.

Assumption 1. For $i = 1, \ldots, n$, $\max\limits_{1 \leq j \leq p} ||x_{ij}/\sqrt{\sigma_{jj}}||_{\psi_2} \leq K^x$ and $||\epsilon_i/\sqrt{\sigma_{\epsilon\epsilon}}||_{\psi_2} \leq K^\epsilon$, where $\sigma_{jj} = \mathrm{Var}(x_{1j}), \sigma_{\epsilon\epsilon} = \mathrm{Var}(\epsilon_1)$.

Assume the indicators are Bernoulli variables with general dependency structure (Dai et al., 2013; Park et al., 2021), that is:

Assumption 2. For $i = 1, \ldots, n$, $(\delta_i^y, \delta_{i1}^x, \ldots, \delta_{ip}^x)$ is from the multivariate Bernoulli distribution with the first two moments written by

$$\mathbb{E}\delta_{ij}^x = \pi_{jj}^{xx}, \quad \mathbb{E}\delta_{ij}^x\delta_i^y = \pi_j^{xy}, \quad \mathbb{E}\delta_{ij}^x\delta_{ik}^x = \pi_{jk}^{xx}.$$

More general moment is denoted as $\mathbb{E}\delta_{ij_1}^x\delta_{ij_2}^x\delta_{ij_3}^x\cdots = \pi_{j_1j_2j_3\ldots}^{xx}$.

The missing mechanism we use is the missing completely at random (MCAR). In the current data structure, we can specify the assumption as follows.

Assumption 3. The data and indicator variables are independent, i.e.

$$\{\epsilon_i, \boldsymbol{x}_i\} \perp\!\!\!\perp \{\delta_i^y, \delta_{i1}^x, \ldots, \delta_{ip}^x\}, \quad i = 1, \ldots, n.$$

The last assumption is about the class of covariance matrices for the covariates. Without loss of generality, assume the variables of interest (i.e. in the set $\mathcal{A}$) are located in front and the covariance matrix $\boldsymbol{\Sigma}$ is decomposed in blocks accordingly.

Assumption 4. Assume the population covariance matrix $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{x}_i)$ satisfies

(a) $\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}$ is positive definite, and

(b) the irrepresentability condition for $\boldsymbol{\Sigma}$ is satisfied with respect to the support set $\mathcal{A}$, i.e., there exists $\tau \in (0, 1)$ such that $\|\boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty \leq 1 - \tau$.

The first condition that the smallest eigenvalue is away from zero is not very restrictive, and the other condition is known to be sufficient and "almost" necessary for selection consistency (van de Geer and Bühlmann, 2009; Lee et al., 2015; Wainwright, 2009).

Throughout this section, we define the LPD estimator as follows. If $\lambda_{\min}(\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}) > 0$, construct the LPD estimator $\Phi_{\mu,\alpha}(\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}})$ by choosing $\alpha = 1$ (and any real-valued μ). Otherwise, for $\epsilon > 0$ such that $\epsilon < \lambda_{\min}(\mathbf{\Sigma})$, set $\alpha = (\mu - \epsilon)/(\mu - \lambda_{\min}(\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}))$ and choose any μ greater than 2ϵ. Based on the assumptions, we present results that guarantee the two conditions (C1) and (C3) in Proposition 1 with high probability.

Theorem 1 (Irrepresentability condition and RSC condition)**.** *Let Assumption 1, 2, 3, 4 hold. Assume $\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}\mathcal{A}}$ is non-singular. Then, the LPD estimator satisfies the irrepresentability condition for some constant $\tilde{\tau} \in (0,1)$ with probability greater than $1 - 3/p^u$ for $u > 0$ if the sample size satisfies*

$$\frac{n}{\pi^{(4)}_{\max} \log p} \geq c \left\{ \frac{\mathrm{tr}(\mathbf{\Sigma}) \max\{(K^x)^2, 1\}\sqrt{u+1}}{\min\{\tau / \left\|\mathbf{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}\right\|_\infty, \lambda_{\min}(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})\}} \right\}^2, \quad n > c\, \pi^{(4)}_{\max} (u+1)^3 \log^3(p \vee n),$$

for some $c > 0$. Here, $\pi^{(4)}_{\max} = \max_{k_1,k_2,\ell_1,\ell_2} \pi^{xx}_{k_1k_2\ell_1\ell_2} / (\pi^{xx}_{k_1\ell_1} \pi^{xx}_{k_2\ell_2})$. Moreover, under the same conditions, (C3) of Proposition 1 holds; if $\lambda_{\min}(\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}) > 0$, μ is excluded in the lower bound of (C3).

To prove the theorem, we first show in Theorem S2 and S3 that the irrepresentability condition holds for $\widehat{\mathbf{\Sigma}}^{\mathrm{LPD}}$ if $\mathbf{\Sigma}$ is in the small neighborhood of the IPW estimator in terms of ℓ_∞, 2-norms. The probability of being in the neighborhood is calculated in the proof of Theorem 1. Technical details can be found in Supplementary Materials S3.1. In Lemma 6 of Datta and Zou (2017), they also showed similar results: if a surrogate estimator $\tilde{\mathbf{\Sigma}}$, which is the LPD estimator in our context, is close enough to $\mathbf{\Sigma}$, then $\tilde{\mathbf{\Sigma}}_{\mathcal{A}^c\mathcal{A}} \tilde{\mathbf{\Sigma}}^{-1}_{\mathcal{A}\mathcal{A}}$ is to $\mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}} \mathbf{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}$. In the theorem below, we use a new notation $\|\boldsymbol{B}\|_{\infty,\mathcal{A}} = \max_{1 \leq j \leq p} \sum_{k \in \mathcal{A}} |b_{jk}|$.

The following guarantees (C2) of Proposition 1 with high probability.

Theorem 2 (Bound on the gradient)**.** *Let Assumption 1, 2, 3 hold. Then, if n and p satisfy*

$$n > c \max\left\{ \log p/\pi^{xy}_{\min}, \pi^{(4)}_{\max} \log^3(p \vee n) \right\}$$

for some $c > 0$, the gradient vector of the mean squared error satisfies the upper bound with probability greater than $1 - 9/p$

$$\left\| \Phi_{\mu,\alpha}(\widehat{\boldsymbol{\Sigma}}^{\mathrm{IPW}})\boldsymbol{\beta}^* - \hat{\boldsymbol{\rho}}^{\mathrm{IPW}} \right\|_\infty \leq L|\mathcal{A}| \sqrt{\frac{\log p}{n}},$$

where $L > 0$ is a function of parameters given by

$$L = C_1 \beta^*_{\max} \max\{(K^x)^2, 1\} \sqrt{\pi^{(4)}_{\max}} \cdot h_1(\mu; \boldsymbol{\Sigma}, \mathcal{A}) + C_2 \frac{\max\left\{ \sqrt{\sigma_{\max}\sigma_{\epsilon\epsilon}} K^x K^\epsilon, \sigma_{\max}(K^x)^2 \right\}}{\sqrt{\pi^{xy}_{\min}}},$$

for some positive constants C_1, C_2. Here, $\pi^{(4)}_{\max} = \max_{k_1,k_2,\ell_1,\ell_2} \pi^{xx}_{k_1k_2\ell_1\ell_2}/(\pi^{xx}_{k_1\ell_1}\pi^{xx}_{k_2\ell_2})$, $\pi^{xy}_{\min} = \min_k \pi^{xy}_k$, $\beta^*_{\max} = \max_{1\leq j\leq p} |\beta^*_j|$, and $h_1(\mu; \boldsymbol{\Sigma}, \mathcal{A}) = \mathrm{tr}(\boldsymbol{\Sigma})\big(1 + \|\boldsymbol{\Sigma}\|_{\infty,\mathcal{A}}/\mu\big)$ if $\lambda_{\min}(\widehat{\boldsymbol{\Sigma}}^{\mathrm{IPW}}) \leq 0$ and $\sigma_{\max}$ otherwise.

Proof of the theorem can be found in Supplementary Materials S3.5. Loh and Wainwright (2013, 2017) also required the bounded gradient condition (see Theorem 1 in Loh and Wainwright (2013) or Loh and Wainwright (2017)). Also, one remarks that dependency of the bound on $\beta^*_{\max}$ is similarly observed in the literature of missing data (see SNR conditions in Chen and Caramanis (2013); Datta and Zou (2017); Theorem 1 in Rosenbaum and Tsybakov (2010)).

Combining these results with Proposition 1, we present the properties of the solution $\widehat{\boldsymbol{\beta}}^{\mathrm{LPD}}$ of (2.7).

Theorem 3. *Let Assumption 1, 2, 3, 4 hold. Assume $\widehat{\boldsymbol{\Sigma}}^{\mathrm{IPW}}_{\mathcal{A}\mathcal{A}}$ is non-singular. We choose the tuning parameter $\lambda \propto L|\mathcal{A}|(\log p/n)^{1/2}$ for the lasso regression. If n and p satisfy*

$$\frac{n}{\pi^{(4)}_{\max} \log p} \geq c \left\{ \frac{\mathrm{tr}(\boldsymbol{\Sigma}) \max\{(K^x)^2, 1\}}{\min\{\tau / \left\|\boldsymbol{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}}\right\|_\infty, \lambda_{\min}(\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}})\}} \right\}^2, \quad n > c \max\left\{ \frac{\log p}{\pi^{xy}_{\min}}, \pi^{(4)}_{\max} \log^3(p \vee n) \right\}$$

for some $c > 0$, *then there exist some* $C > 0, d > 0, \tilde{\tau} \in (0,1)$ *such that we can guarantee with probability greater than* $1 - d/p$

$$(R1) \quad \textit{The minimizer } \widehat{\boldsymbol{\beta}}^{\mathrm{LPD}} \textit{ is unique.}$$

$$(R2) \quad \|\widehat{\boldsymbol{\beta}}^{\mathrm{LPD}} - \boldsymbol{\beta}^*\|_2 \leq C \frac{L}{\tilde{\tau} \cdot h_2(\mu, \lambda_{\min}(\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}))} \sqrt{\frac{|\mathcal{A}|^3 \log p}{n}}$$

$$(R3) \quad \hat{\beta}_j^{\mathrm{LPD}} = 0, \quad j \in \mathcal{A}^c.$$

Here, $h_2(\mu, \lambda_{\min}(\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}})) = \min\{\lambda_{\min}(\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}), \mu\}$ *if* $\lambda_{\min}(\widehat{\boldsymbol{\Sigma}}^{\mathrm{IPW}}) \leq 0$ *and* $\lambda_{\min}(\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}})$ *otherwise. The factor* L *appears in Theorem 2.*

We have some remarks regarding this main result. First, the results hold regardless of the choice of matrix norms in (2.6) because the optimal choice of α in LPD is independent of the matrix norms. Also, no terms are involved with ϵ in the theorems, though the actual performance of LPD can change according to different ϵ due to the numerical stability.

Second, the constant L depends on $\mathrm{tr}(\boldsymbol{\Sigma})$, which is an order of p in general. This trace term is introduced when we control the magnitude of the gradient vector of the loss function based on the LPD. This condition related to the gradient vector is commonly used in literature (e.g. (3.1) of Loh and Wainwright (2012)). We believe that the additional factor is the expense in theory we need to pay for convexification of the loss function. However, the empirical performance of the propose method scales with $\sqrt{\log p/n}$ up to an multiplicative constant not depending on n and p, which is presented in Supplementary Material S4.5.

Moreover, there exists a class of distributions of covariates where the constant L is independent of p. As in the literature on covariance estimation (Lounici (2014); Mendelson and Zhivotovskiy (2020); Koltchinskii and Lounici (2017)), we can express the trace of $\boldsymbol{\Sigma}$ by the effective rank that measures intrinsic dimension of a symmetric matrix, defined by

$\mathbf{r}(\mathbf{\Sigma}) = \mathrm{tr}(\mathbf{\Sigma})/||\mathbf{\Sigma}||_2$. Note that $\mathbf{r}(\mathbf{\Sigma}) \leq \mathrm{rank}(\mathbf{\Sigma}) \leq p$ for general matrices, but the effective rank would be much smaller than p if $\mathbf{\Sigma}$ is approximately low-rank. See more discussion in Section 2.2 of Lounici (2014) or Remark 5.53 of Vershynin (2011). Hence, the constant L would not depend on p if we consider a class of covariance matrices satisfying that (1) approximately low-rank, or $\mathrm{r}(\mathbf{\Sigma}) := \mathrm{tr}(\mathbf{\Sigma})/||\mathbf{\Sigma}||_2 \leq R$ (independent of p) and (2) the largest eigenvalue is bounded, or $||\mathbf{\Sigma}||_2 \leq B$ (independent of p). Note that the boundedness of a trace of low-rank matrix Σ does not contradict to Assumption 4. Then, Theorem 3 states that under this class of distributions for covariates, the sample size $n \gtrsim \log p$ is enough to guarantee that the solution $\widehat{\boldsymbol{\beta}}^{\mathrm{LPD}}$ is (R1) unique, (R2) ℓ_2-consistent, and (R3) has no false positive with probability close to 1.

Third, we would like to compare our result with the ones previously obtained in Datta and Zou (2017) and Loh and Wainwright (2012). To facilitate a fair comparison, we reorganize all the results into the following format: if the sample size and dimension satisfies $n/\log p > \mathcal{M}$, then with probability at least $1 - c/p$, it holds that

$$||\widehat{\beta} - \beta^*||_2 \leq C \cdot \mathcal{L} \cdot |\mathcal{A}|^{\mathcal{K}} \sqrt{\frac{\log p}{n}},$$

where $c, C > 0$ are some positive constants. Here, $\widehat{\beta}$ is a coefficient estimator from one of Datta and Zou (2017), Loh and Wainwright (2012), or the proposed, and β^* is the true value to be estimated. The specific forms of $\mathcal{K}$, $\mathcal{L}$, and $\mathcal{M}$ depend on parameters such as (but not limited to) (1) observation probability, (2) tail thickness (or sub-Gaussian parameter) of the response variable, (3) tail thickness of the covariates, (4) covariance matrix of the covariates. While the triplet $(\mathcal{K}, \mathcal{L}, \mathcal{M})$ is not directly comparable as each paper uses slightly different assumptions, we aim to highlight the general tendencies.

The convergence rate $\mathcal{L}$ commonly depends on (1) observation probability, (2) tail thick-

ness (or sub-Gaussian parameter) of the response variable, (3) tail thickness of the covariates, (4) magnitude of the true value β^*, and (5) well-conditionedness of $\mathbf{\Sigma}$. Regarding (5), the result from Loh and Wainwright (2012) is $\mathcal{L} \propto 1/\lambda_{\min}(\mathbf{\Sigma})$, while Datta and Zou (2017) obtained $\mathcal{L} \propto 1/\Omega$, where

$$\Omega := \min_{x \in \mathcal{R}} x^\top \mathbf{\Sigma} x, \qquad \mathcal{R} = \{x : ||x||_2 = 1, ||x_{\mathcal{A}^c}||_1 \leq 3||x_{\mathcal{A}}||_1\},$$

which is related to the compatibility condition. In contrast, our result satisfies $\mathcal{L} \propto 1/\{\tilde{\tau} \cdot (\lambda_{\min}(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}}) \wedge \mu)\}$, where $\tilde{\tau}$ is a constant from the irrepresentability condition of the LPD estimator. Similar quantities have appeared from restricted strong convexity in the related context (Negahban et al. (2012)), typically with the same order of 1 in the denominator. The rate from Loh and Wainwright (2012) would get worse if the covariance matrix from covariates on $\mathcal{A}^c$ is ill-conditioned, while the other two are not affected. Additionally, while our result depends on μ (the tuning parameter of LPD procedure), this dependency is negligible if μ is chosen sufficiently large, i.e., $\mu > \lambda_{\min}(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})$. Lastly, our result has dependency on $\mathrm{tr}(\mathbf{\Sigma})$, i.e. $\mathcal{L} \propto \mathrm{tr}(\mathbf{\Sigma})$.

The constant $\mathcal{M}$ characterizes the sample size required to guarantee the derived convergence rate. Across all three methods, the constant depends on (1) observation probability, (2) tail thickness of the covariates, and (3) well-conditionedness of $\mathbf{\Sigma}$. The dependency on (3) is similar to that of $\mathcal{L}$. More specifically,

$$\mathcal{M}_{\text{Loh}} \propto 1/\lambda_{\min}(\mathbf{\Sigma})^2, \quad \mathcal{M}_{\text{Datta}} \propto 1/\min\{C_1\tau^2, C_2\Omega^2\}, \quad \mathcal{M}_{\text{Park}} \propto 1/\{\tau \cdot \lambda_{\min}(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})\}^2$$

where $C_1, C_2 > 0$ are constants. In Datta and Zou (2017), $\mathcal{M}$ also depends on $\beta^*_{\max}$ and the tail thickness of the response variable. In our case, $\mathcal{M} \propto \mathrm{tr}(\mathbf{\Sigma})$, which can be explained similarly to its appearance in $\mathcal{L}$.

The constant $\mathcal{K}$ represents the order of sparsity in the convergence rate. Both Datta and Zou (2017) and our result share the same order $\mathcal{K} = 3/2$, while Loh and Wainwright (2012) achieves a smaller order $\mathcal{K} = 1$. The order of sparsity may have room for improvement in proof techniques, as the exponent $\mathcal{K} = 1/2$ in $|\mathcal{A}|^{\mathcal{K}}$ is commonly observed in the high-dimensional regression literature (e.g. van de Geer and Bühlmann (2009); Wainwright (2009); Negahban et al. (2012)). In contrast, our result yields $\mathcal{K} = 3/2$, which is attributed to the linear shrinkage of the non-PD matrix. This can also be seen as a cost incurred for convexification.

In conclusion, this comparison shows that our method still guarantees similar results from the previous work, but with an extra term $\text{tr}(\mathbf{\Sigma})$. Theoretically, this difference is the price we need to pay for convexification and faster computation. However, for a smaller class of covariance matrices (e.g., low-rank and bounded largest eigenvalue), this term becomes negligible.

2.4 Estimation of unknown parameters

It should be noted that our results are based on two implicit assumptions. First, we assume the observation probabilities are known, as in other error-in-variable literatures (Datta and Zou (2017); Sørensen et al. (2015)). Second, following a convention in a regression framework, we also assume covariates are centered, i.e. mean-zero. However, these may not be the case in real-world data, and thus we would like to leave some remarks regarding these assumptions.

For estimating the observation probabilities, it is natural to use the empirical proportions (i.e. the proportion of observed pairs) under MCAR, due to the law of large numbers. In other words, we suggest using $\hat{\pi}^{xx}_{jk} = \sum_{i=1}^{n} \delta^x_{ij}\delta^x_{ik}/n$ and $\hat{\pi}^{xy}_{j} = \sum_{i=1}^{n} \delta^x_{ij}\delta^y_{i}/n$. Then, the new

IPW estimator is

$$\widehat{\mathbf{\Sigma}}^{\mathrm{IPW},\hat{\pi}} = \Big((\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}})_{jk} \frac{\pi^{xx}_{jk}}{\hat{\pi}^{xx}_{jk}},\ 1 \le j,k \le p \Big).$$

We have found throughout our numerical study that the penalized regression based on the above estimator performs quite well.

Next, we consider the case when covariates may have non-zero means. The most straightforward way is to center each covariate by the IPW mean estimator $\hat{\mu}_j = \frac{\sum_{i=1}^n \tilde{x}_{ij}}{n\pi^{xx}_{jj}}$. As used in Kolar and Xing (2012) and Cai and Zhang (2016), this type of IPW estimator is defined by $\widehat{\mathbf{\Sigma}}^{\mathrm{IPW},2}_{jk} = \sum_{i=1}^n \delta^x_{ij}\delta^x_{ik}(\tilde{x}_{ij} - \hat{\mu}_j)(\tilde{x}_{ik} - \hat{\mu}_k)/(n\pi^{xx}_{jk})$. However, this is not unbiased (in finite sample), which often complicates theoretical analyses (e.g. concentration inequality). To address it, we proposed another type of IPW estimator in our earlier work (Park et al. (2021)):

$$\widehat{\mathbf{\Sigma}}^{\mathrm{IPW},3}_{jk} = \frac{\sum_{i=1}^n \tilde{x}_{ij}\tilde{x}_{ik}}{n\pi^{xx}_{jk}} - \frac{\sum_{i\neq i'}^n \tilde{x}_{ij}\tilde{x}_{i'k}}{n(n-1)\pi^{xx}_{jj}\pi^{xx}_{kk}}.$$

We remark that our theory is based on two types of concentration inequalities for IPW estimators: one is about the element-wise maximum norm and the other is the spectral norm. The former has been investigated in our earlier work (Park et al. (2023)), but the latter has not yet in literature. Though we tried to derive the non-asymptotic inequality based on the spectral norm, it is not as simple as the other. We think including such an analysis in this paper would be unnecessarily complicated, and thus leave it as our future work.

3. Numerical study

We showcase the empirical performance of the proposed estimator LPD based on different simulation parameters (e.g. dimension p, missing rate of observations, covariance structure for variables). Our analysis consists of three parts. In the first part, we compare several

methods including two existing ones and the proposed one based on different choices of μ. In the second, we examine how sensitive the models are to missing values. In the third, we time an algorithm of each method to see their scalability.

It has to be noted that a simulation study performed by Romeo and Thoresen (2019) compared a group of methods available until then, but only considered additive measurement error models. In the meantime, our simulation study deals with missing data cases, which is clearly different from what was covered in their work.

3.1 Setting

We adopt experimental settings of Sørensen (2019) where they generate responses from the normal model, i.e. $\tilde{\boldsymbol{y}} \sim N_n(\tilde{\boldsymbol{X}}\boldsymbol{\beta}^*, \sigma_y^2\mathbf{I})$, and each row of the design matrix $\tilde{\boldsymbol{X}}$ from $N(\mathbf{0}, \boldsymbol{\Sigma})$ where the covariance structure is the compound symmetry ($\boldsymbol{\Sigma}_{ij} = 0.5^{\mathrm{I}(i\neq j)}$). The dimension p of covariates varies over $p = 200, 500$. The regression coefficients $\boldsymbol{\beta}^*$ have non-zero values at random positions while keeping the proportion of them at $s = 0.05, 0.1$ (i.e. s is the level of sparsity). The non-zero coefficients are all equal to 1. We fix $n = 200$ and $\sigma_y = 3$.

Responses and covariates are subject to missing completely at random (MCAR). More specifically, we define matrices of missing indicators: $\boldsymbol{M}_y = (\delta_i^y)$ and $\boldsymbol{M}_X = (\delta_{ij}^x)$ where $\delta_i^y \sim \mathrm{Ber}(\theta)$, $\delta_{i,3j}^x \sim \mathrm{Ber}(\theta)$, $j = 1, \ldots, \lfloor p/3 \rfloor$, independently. Then, the corrupted data are

$$\boldsymbol{y} = \tilde{\boldsymbol{y}} * \boldsymbol{M}_y, \quad \boldsymbol{X} = \tilde{\boldsymbol{X}} * \boldsymbol{M}_X,$$

where $*$ is the element-wise product. Other missing mechanisms (MAR, MNAR) will be discussed in Section S4.3. We control the observation probability $\theta = 0.7, 0.9$. We generate 100 independent datasets to consider random variability.

Given incomplete data $(\boldsymbol{y}, \boldsymbol{X})$, we compute three comparative estimators: (1) linear

shrinkage positive definite lasso (LPD), (2) convex conditioned lasso (CoCo) (Datta and Zou, 2017), and (3) non-convex lasso (NCL) (Loh and Wainwright, 2012). We use the R package named `BDcocolasso` (available at `https://github.com/celiaescribe/BDcocolasso`) implemented by Escribe et al. (2021) to obtain the second estimator and `hdme` (Sørensen, 2019) to obtain the third. Additionally, we add two types of lasso regression in comparison. One uses the complete data $(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{X}})$ and is named (4) "true lasso", while the other runs the lasso regression with mean imputed data and is named (5) "naive lasso". We do not include the complete-case analysis as none of the samples are completely observed in high-dimensional missing data. For instance, in the real data we analyzed, every cell line has at least 48 missing values, making the straightforward approach impractical.

In terms of LPD, we can consider a set of variants based on different choices of μ, but found that LPD using ℓ_∞-norm empirically works well and is robust to different setups. Hence, for readability, we only report the corresponding results in this section, while the entire results are provided in Supplementary Materials S4.2 and S4.3.

The penalized regression methods mentioned earlier have hyperparameters to be tuned. To choose a penalty parameter λ of CoCo and LPD, we use the corrected cross-validation proposed in Datta and Zou (2017), that is, the cross-validation approach adjusted for corrupted data. Simply put, the idea is to minimize the mean square prediction error where a non-PD covariance matrix estimate is replaced by the PD matrix. More details can be found in Supplementary Materials S4.1. The grids are evenly spaced in log scale within the interval $[R/10000, R]$ where $R = 2||\boldsymbol{r}_{\text{naive}}||_{\max}$ and $\boldsymbol{r}_{\text{naive}}$ is the naive lasso estimator. If $R = 0$ (i.e. $\boldsymbol{r}_{\text{naive}} = 0$), then we set R by $||\boldsymbol{X}^\top \boldsymbol{y}/n||_{\max}$. For NCL, we need to decide the radius b such that the solution satisfies $||\widehat{\boldsymbol{\beta}}||_1 \leq b$. We search the optimal radius over the grid in

$[R/10000, R]$ with $R = 2||\boldsymbol{r}_{\text{naive}}||_1$. The number of grid points is 100 throughout. Using the optimal tuning parameter, we re-fit each model and have the estimates of coefficients.

We measure six criteria to assess performance of each method. Following Datta and Zou (2017), we compute the prediction error (PE) and mean squared error (MSE), which is respectively defined

$$PE(\widehat{\boldsymbol{\beta}}) = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\top \boldsymbol{\Sigma} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*), \quad MSE(\widehat{\boldsymbol{\beta}}) = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\top (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*).$$

The number of covariates corrected/incorrectly identified (TP and FP) are also counted. To see an overall accuracy of variable selection, we also compute the (partial) area under the ROC curve (pAUC) and F_1-score (harmonic mean of precision and recall) denoted by F1. We also measure the time each method would take to finish. This includes the tuning parameter search.

3.2 Results

We present the results of our numerical study, which consists of three parts: (1) a comparison of different regression methods, (2) a sensitivity analysis with respect to missing rates and missing mechanisms, and (3) a timing analysis for computation. Due to space limitations, the latter two are provided in Supplementary Material (see Section S4.3 and S4.4).

In the followings, we compare different regression methods. To reduce the workload of simulations, we fix $\theta = 0.9$ under MCAR. Compared with the existing methods (CoCo, NCL), LPD is less sparser and has more TP and FP. LPD is proved to be successful in estimation (low MSE), prediction (low PE), and variable selection (high pAUC, high TP). Though the difference is negligible considering standard deviation, LPD performs best in almost all scenarios of the finite sample setting. This result is of great importance since

	$p = 200, s = 0.05$					
	PE	MSE	pAUC	F_1	TP	FP
TL	1.892 (0.601)	3.653 (1.162)	0.953 (0.032)	0.439 (0.065)	9.700 (0.482)	25.370 (6.935)
NL	3.710 (1.279)	6.186 (1.950)	0.873 (0.075)	0.397 (0.076)	8.560 (1.157)	25.590 (7.732)
CoCo	3.490 (1.276)	6.641 (2.424)	0.816 (0.073)	0.398 (0.083)	8.370 (1.236)	24.650 (6.658)
NCL	5.162 (1.337)	6.447 (1.820)	0.519 (0.083)	0.439 (0.118)	8.140 (1.477)	21.800 (15.525)
LPD	3.352 (1.000)	6.320 (1.824)	0.873 (0.070)	0.369 (0.066)	8.790 (1.104)	29.710 (7.312)
	$p = 500, s = 0.05$					
	PE	MSE	pAUC	F_1	TP	FP
TL	6.073 (1.243)	11.940 (2.433)	0.815 (0.044)	0.420 (0.054)	22.980 (1.239)	63.190 (16.677)
NL	16.327 (4.124)	26.382 (4.161)	0.555 (0.084)	0.298 (0.060)	13.130 (3.084)	49.950 (9.090)
CoCo	15.738 (3.154)	30.083 (5.651)	0.600 (0.044)	0.290 (0.062)	12.530 (3.119)	48.810 (9.018)
NCL	27.640 (7.481)	26.873 (3.507)	0.506 (0.062)	0.218 (0.055)	14.810 (5.025)	105.450 (55.242)
LPD	13.375 (2.323)	25.482 (3.883)	0.717 (0.064)	0.262 (0.050)	15.250 (3.141)	76.730 (16.213)
	$p = 200, s = 0.1$					
	PE	MSE	pAUC	F_1	TP	FP
TL	3.240 (0.841)	6.263 (1.631)	0.915 (0.034)	0.535 (0.060)	19.600 (0.651)	34.570 (8.335)
NL	10.299 (3.229)	15.240 (3.293)	0.761 (0.068)	0.438 (0.062)	14.400 (2.340)	31.500 (5.458)
CoCo	9.361 (2.429)	17.288 (4.059)	0.723 (0.055)	0.437 (0.070)	13.880 (2.341)	29.950 (6.660)
NCL	16.726 (3.676)	17.447 (2.445)	0.617 (0.046)	0.398 (0.099)	14.170 (2.775)	42.950 (26.712)
LPD	8.477 (2.144)	15.565 (3.406)	0.774 (0.060)	0.419 (0.057)	14.970 (2.115)	36.940 (7.678)
	$p = 500, s = 0.1$					
	PE	MSE	pAUC	F_1	TP	FP
TL	14.001 (2.440)	27.630 (4.914)	0.683 (0.049)	0.477 (0.048)	43.950 (2.488)	91.930 (18.908)
NL	48.644 (11.035)	77.535 (11.147)	0.391 (0.057)	0.269 (0.055)	16.770 (3.928)	57.530 (9.157)
CoCo	47.577 (8.028)	91.880 (15.888)	0.548 (0.033)	0.259 (0.051)	15.560 (3.529)	54.000 (8.060)
NCL	76.542 (26.472)	65.129 (11.035)	0.489 (0.039)	0.241 (0.036)	24.940 (7.538)	129.610 (44.213)
LPD	37.225 (5.155)	71.559 (9.319)	0.606 (0.043)	0.267 (0.045)	21.020 (4.259)	86.310 (15.103)

Table 1: Method comparison for $p = 200, 500$ and $s = 0.05, 0.1$. Each performance measure is averaged over $R = 100$ repetitions (standard deviation in parenthesis).

LPD is much faster than its competitors (see Table S4). The naive lasso (NL) seems to have smaller MSE and higher F_1-score than LPD, but it sharply deteriorates when p increases. Compared to it, LPD performs nearly best for all cases considered.

Though its more restrictive structure in LPD than CoCo, it shows the superior performance in the finite sample study. We believe this is because LPD preserves the off-diagonal elements of the initial estimator. That is, LPD does not change information about the co-

variance part. In contrast, CoCo focuses on element-wise approximation, which may lose such information. As a result, CoCo has good theoretical support, but LPD offers a more practical solution.

4. Real data: Genomics of Drug Sensitivity in Cancer

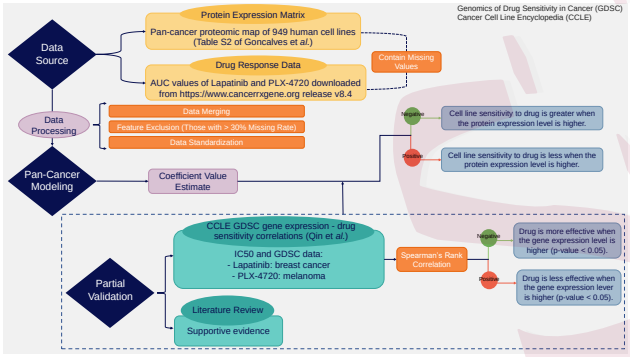

Figure 1: The overview of the pan-cancer drug sensitivity analysis and partial validation.

In this section, we studied the performance of the proposed method through drug response data available from Genomics of Drug Sensitivity in Cancer (GDSC). In this dataset, cancer cell lines (samples) are treated with different drugs or compounds. Sensitivity to some drugs was measured by the area under the doseresponse curve (AUC_{RS}) (a response variable), which is to be modeled by the protein levels of cells (explanatory variables). A small AUC_{RS} value indicates a strong drug response of the cell line to the drug. A large value of AUC_{RS} means no or limited response of the cell line to the tested drug (Vis et al., 2016). Among many, we used the protein expression data from 949 human cancer cell lines. We aimed to discover a list of (small portion of) proteins (biomarkers) that help explain the drug sensitivity for the anti-cancer drug of interest. These lists may also be used to identify cell lines that respond to some drugs more actively than others.

In the dataset, 949 cell lines and 8,498 protein expressions were incompletely measured, but we deleted proteins in which more than 30% of values were missing, resulting in the bottom left of Figure 2. Then, the final data we used to analyze is $n = 867$ cell lines and $p = 4,183$ proteins. It has 7.15% of missing values in average across cell lines (see the top of Figure 2). However, every cell line has at least 48 missing values (see the bottom right of Figure 2), meaning the listwise deletion is not feasible.

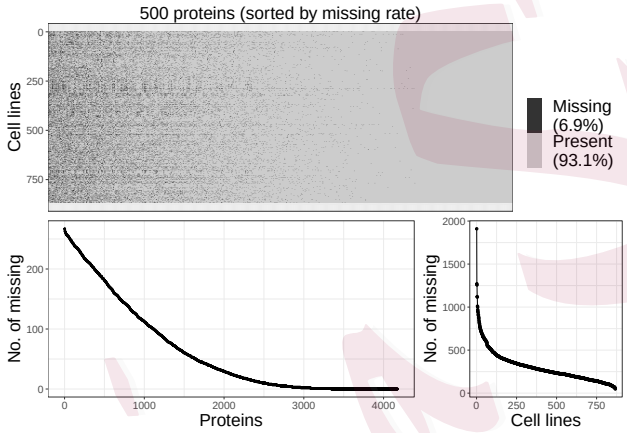

Figure 2: In the top figure, missing values are marked as black in the data matrix with randomly chosen 500 proteins. The two bottom figures show the number of missing values in either proteins (left) or cell lines (right).

We used Lapatinib (an approved drug in treating HER2-positive breast cancers, an inhibitor of EGFR (also known as ERBB1 and HER1) (Xu et al., 2017) and HER2 (also known as ERBB2)) and PLX-4720 (selective inhibitor of BRAFV600E) as two examples to showcase the application of our method in examining the pan-cancer drug responses and exploring potential protein biomarkers of cancer vulnerabilities.

Before running our proposed method based on ℓ_∞-norm, we standardized AUC_{RS} and protein expressions using sample means and standard deviations calculated ignoring missing

values. The grid search for the tuning parameter was similarly performed as in the simulation study; the naive lasso estimator $\boldsymbol{r}_{\text{naive}}$ was fit and used to decide the range of grids $[R/10000, R]$ with $R = 2||\boldsymbol{r}_{\text{naive}}||_{\max}$ in which 100 evenly spaced grid points were considered. The cross-validation error curves are given in the left of Figure 3.

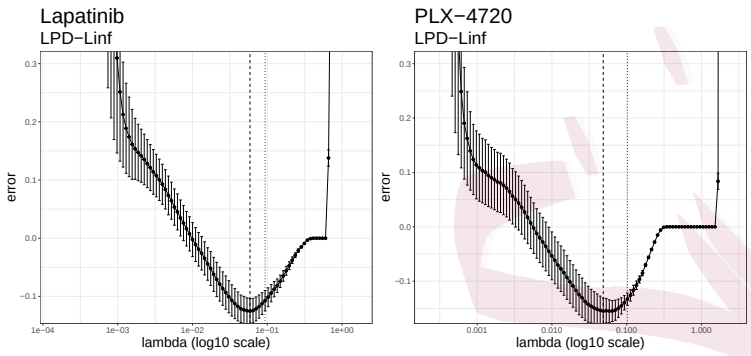

Figure 3: The corrected cross-validation error (solid line). The two vertical lines indicate the optimal tuning parameter (dashed line) and 1-se rule (dotted line), respectively. The error bar is deviated from the center by one standard error.

We attempted to interpret the estimated coefficients. For simplicity, we applied the 1-se rule (the dotted line in Figure 3) that chose a slightly larger tuning parameter and pursued a sparser solution whose accuracy was still acceptable. Table 2 below shows the number of non-zero coefficients and their signs.

	Lapatinib			PLX-4720		
Sign	(−)	(+)	zero	(−)	(+)	zero
Count	48	40	4088	58	29	4089

Table 2: Signs of the estimated coefficients from the 1-se rule (transposed).

In our analysis, a negative association (coefficient) with AUC_{RS} suggests greater sensitivity

(of a cell line) when the protein level is high. A tool developed by Qin et al. (2017) aiming at the discovery of drug sensitivity and gene expression association was used to assist us in demonstrating the robustness of our method. In Qin et al. (2017), a positive correlation with the IC50 indicates that the drug is less effective when the expression of a targeted gene is higher and vice versa. However, it is essential to note that the concordance between proteomics and transcriptomics can be weak (Wu et al., 2013). Integrating the information obtained from each data modality may help predict the effects of gene/protein levels on anti-cancer drug activity (Gonçalves et al., 2022).

For the case of Lapatinib, we found 48 proteins that showed a significant negative association with the AUC_{RS}. Interestingly, EGFR, the canonical target of Lapatinib, was also found to be among the selected proteins. Among 48 proteins, nine showed concordance with the expression of nine genes (*BAIAP2*, *FAM83H*, *HDHD3*, *HSD17B8*, *KRT19*, *MIEN1*, *PLXNB2*, *REEP6*, and *SEC16A*) affecting the activity of Lapatinib estimated by Qin et al. (2017) using IC50 and GDSC gene expression data. It has been known that *MIEN1* is amplified along *ERBB2* and exhibits oncogenic potential (Omenn et al., 2014). It is linked to cisplatin resistance and is highly expressed in Lapatinib-sensitive breast cancer cells than Lapatinib-resistant breast cancer cells (Kumar et al., 2019).

PLX-4720 has shown *in vitro* and *in vivo* efficacy in treating thyroid cancer and melanoma (Coperchini et al., 2019). In our analysis, 58 proteins showed a negative association with AUC_{RS}. Regarding thyroid cancer, 8 corresponding genes (*FAHD2A*, *FKBP10*, *GSN*, *QDPR*, *RAB27A*, *RETSAT*, *S100A13*, *TIMM50*) also had negative Spearman's rank correlation coefficient in the analysis by Qin et al. (2017) (using IC50 and GDSC gene expression data). Ten out of 12 genes (*AMDHD2*, *CTSB*, *ENDOD1*, *HIBADH*, *KANK2*, *PML*, *RPS27L*, *SP100*,

STX7, and *TIMMDC1*) showed negative Spearman's rank correlation coefficient in the analysis for melanoma by Qin et al. (2017). These generally concordant results suggest the relevance of our pan-cancer regression modeling approach.

5. Conclusion

This paper tackles the penalized linear regression problem with missing observations where the estimated Gram matrix of covariates is non-PD in general. To handle it, we present a significantly simpler approach for positive definite modification of non-PD matrices inspired by linear shrinkage of covariance matrix. Due to its closed forms, the procedure is scalable even for high-dimensional regression, while the lasso solution based on it still enjoys the same rate of convergence and selection consistency. Through analyzing simulated and real data, we verify that the proposed method has a greater advantage in computational aspect compared to existing methods while ensuring theoretical properties such as selection consistency.

We acknowledged some potential to extend our method to the MAR case by modeling the observation probability $\pi^{xx}_{i,jk} = \pi(\boldsymbol{x}_{i,\mathrm{obs}}; \boldsymbol{\eta})$ using the (fully) observed data. It can be shown that the corresponding IPW estimator is unbiased under the MAR assumption, but its concentration inequalities are more difficult to derive due to the dependency of observed data. This extension is interesting for future work. Moreover, we expressed the estimation performance with the minimum pairwise sample size. Zheng and Allen (2023) came up with measuring individual dependency on missing observations in a different context (estimation of the graphical model). Under suitable assumptions on the graph structure of explanatory variables (e.g. sparsity), representing the individual dependency would give more insights for the regression coefficients. This needs more investigation on the simultaneous estimation

of covariance matrix and regression coefficients, and thus we leave it as future work.

As the quadratic loss is closely connected to the Gaussian distribution, a natural extension of our work is to exponential families, i.e. the generalized linear model (GLM). Seemingly, it looks challenging to define a Gram matrix in this context due to the non-linear link function. However, when fitting the generalized linear model, an adjusted dependent variable is used in the process of an iterative (re-)weighted least squares (James and Radchenko (2009)). Moreover, one may find that the adjusted dependent variable can be seen as the sum of a linear predictor (evaluated at the current iteration) and the Pearson residual. Based on this observation, we may construct Gram matrices defined between linear predictors and/or Pearson residuals. We plan to explore this extension in future.

To address the sub-optimal convergence rate caused by the trace term in our theories, there might be room for improvement. Currently, we transit the deviation of the smallest eigenvalue of the IPW estimator (see Lemma S3) to the spectral norm using Weyl's inequality; $|\lambda_{\min}(\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}}) - \lambda_{\min}(\mathbf{\Sigma})| \leq ||\widehat{\mathbf{\Sigma}}^{\mathrm{IPW}} - \mathbf{\Sigma}||_2$. However, this inequality may not be tight in a certain class $\tilde{\mathcal{C}}$ of the covariance matrix. If a sharper upper bound of the left-hand side, ideally not depending on the trace term, could be achieved, then the theoretical results could be further improved.

Supplementary Materials

The supplementary material available online presents additional simulation results and technical theorems to prove the main results.

Acknowledgements

We thank the anonymous reviewers for their constructive comments and valuable suggestions, which have greatly improved the quality of this work.

Seongoh Park was supported by the government of the Republic of Korea (MSIT) and the National Research Foundation of Korea (NRF-1711200203); the Sungshin Womens University Research Grant of H20240073. Johan Lim was supported by the government of the Republic of Korea (MSIT) and the National Research Foundation of Korea (NRF-2021R1A2C1010786)

References

Bach, F. R. (2008). Bolasso: Model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, New York, NY, USA, pp. 3340. Association for Computing Machinery.

Bickel, P. J. and E. Levina (2008a). Covariance regularization by thresholding. *The Annals of Statistics 36*(6), 2577 – 2604.

Bickel, P. J. and E. Levina (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics 36*(1), 199 – 227.

Bien, J. and R. J. Tibshirani (2011, 12). Sparse estimation of a covariance matrix. *Biometrika 98*(4), 807–820.

Cai, T. T. and A. Zhang (2016). Minimax rate-optimal estimation of high-dimensional covariance matrices with incomplete data. *Journal of Multivariate Analysis 150*, 55–74.

Candes, E. and T. Tao (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics 35*(6), 2313 – 2351.

Chatterjee, A. and S. N. Lahiri (2011). Bootstrapping lasso estimators. *Journal of the American Statistical Associa-*

tion 106(494), 608–625.

Chatterjee, A. and S. N. Lahiri (2013). Rates of convergence of the Adaptive LASSO estimators to the Oracle distribution and higher order refinements by the bootstrap. *The Annals of Statistics 41*(3), 1232 – 1259.

Chen, Y. and C. Caramanis (2013, 17–19 Jun). Noisy and missing data regression: Distribution-oblivious support recovery. In S. Dasgupta and D. McAllester (Eds.), *Proceedings of the 30th International Conference on Machine Learning*, Volume 28 of *Proceedings of Machine Learning Research*, Atlanta, Georgia, USA, pp. 383–391. PMLR.

Cho, S., S. Katayama, J. Lim, and Y.-G. Choi (2021, Dec). Positive-definite modification of a covariance matrix by minimizing the matrix ℓ_∞norm with applications to portfolio optimization. *AStA Advances in Statistical Analysis 105*(4), 601–627.

Choi, Y.-G., J. Lim, A. Roy, and J. Park (2019). Fixed support positive-definite modification of covariance matrix estimators via linear shrinkage. *Journal of Multivariate Analysis 171*, 234–249.

Coperchini, F., L. Croce, M. Denegri, O. Awwad, S. T. Ngnitejeu, M. Muzza, V. Capelli, F. Latrofa, L. Persani, L. Chiovato, and M. Rotondi (2019, Mar). The braf-inhibitor plx4720 inhibits cxcl8 secretion in brafv600e mutated and normal thyroid cells: a further anti-cancer effect of braf-inhibitors. *Scientific Reports 9*(1), 4390.

Dabke, K., S. Kreimer, M. R. Jones, and S. J. Parker (2021). A simple optimization workflow to enable precise and accurate imputation of missing values in proteomic data sets. *Journal of Proteome Research 20*(6), 3214–3229. PMID: 33939434.

Dai, B., S. Ding, and G. Wahba (2013, 09). Multivariate bernoulli distribution. *Bernoulli 19*(4), 1465–1483.

Datta, A. and H. Zou (2017). CoCoLasso for high-dimensional error-in-variables regression. *The Annals of Statistics 45*(6), 2400 – 2426.

Daye, Z. J., J. Chen, and H. Li (2012). High-dimensional heteroscedastic regression with an application to eqtl data analysis. *Biometrics 68*(1), 316–326.

Du, J., J. Boss, P. Han, L. J. Beesley, M. Kleinsasser, S. A. Goutman, S. Batterman, E. L. Feldman, and B. Mukherjee

(2022). Variable selection with multiply-imputed datasets: Choosing between stacked and grouped methods. *Journal of Computational and Graphical Statistics 31*(4), 1063–1075.

Duchi, J. and Y. Singer (2009). Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research 10*(99), 2899–2934.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of Statistics 32*(2), 407 – 499.

Escribe, C., T. Lu, J. Keller-Baruch, V. Forgetta, B. Xiao, J. B. Richards, S. Bhatnagar, K. Oualkacha, and C. M. T. Greenwood (2021). Block coordinate descent algorithm improves variable selection and estimation in error-in-variables regression. *Genetic Epidemiology 45*(8), 874–890.

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.

Friedman, J., T. Hastie, H. Höfling, and R. Tibshirani (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics 1*(2), 302 – 332.

Fu, W. and K. Knight (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics 28*(5), 1356 – 1378.

Ghosh, D. and A. M. Chinnaiyan (2005). Classification and selection of biomarkers in genomic data using lasso. *Journal of Biomedicine and Biotechnology 2005*, 147–154.

Gonçalves, E., R. C. Poulos, Z. Cai, S. Barthorpe, S. S. Manda, N. Lucas, A. Beck, D. Bucio-Noble, M. Dausmann, C. Hall, M. Hecker, J. Koh, H. Lightfoot, S. Mahboob, I. Mali, J. Morris, L. Richardson, A. J. Seneviratne, R. Shepherd, E. Sykes, F. Thomas, S. Valentini, S. G. Williams, Y. Wu, D. Xavier, K. L. MacKenzie, P. G. Hains, B. Tully, P. J. Robinson, Q. Zhong, M. J. Garnett, and R. R. Reddel (2022). Pan-cancer proteomic map of 949 human cell lines. *Cancer Cell 40*(8), 835–849.e8.

Han, F., J. Lu, and H. Liu (2014). Robust scatter matrix estimation for high dimensional distributions with heavy tails. *Technical report, Princeton University*.

Han, Y. and R. S. Tsay (2020). High-dimensional linear regression for dependent data with applications to nowcasting. *Statistica Sinica 30*(4), 1797–1827.

Heymans, M. W., S. van Buuren, D. L. Knol, W. van Mechelen, and H. C. de Vet (2007, Jul). Variable selection under multiple imputation using the bootstrap in a prognostic study. *BMC Medical Research Methodology 7*(1), 33.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*(1), 55–67.

James, G. M. and P. Radchenko (2009, 06). A generalized Dantzig selector with shrinkage tuning. *Biometrika 96*(2), 323–337.

Karpievitch, Y. V., A. R. Dabney, and R. D. Smith (2012, Nov). Normalization and missing value imputation for label-free lc-ms analysis. *BMC Bioinformatics 13*(16), S5.

Kolar, M. and E. P. Xing (2012). Estimating sparse precision matrices from data with missing values. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, USA, pp. 635–642. Omnipress.

Koltchinskii, V. and K. Lounici (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli 23*(1), 110 – 133.

Kumar, S., P. P. Kushwaha, and S. Gupta (2019). Emerging targets in cancer drug resistance. *Cancer Drug Resistance 2*(2), 161–177.

Lachenbruch, P. A. (2011). Variable selection when missing values are present: a case study. *Statistical Methods in Medical Research 20*(4), 429–444. PMID: 20442196.

Lam, C. and J. Fan (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics 37*(6B), 4254 – 4278.

Langford, J., L. Li, and T. Zhang (2008). Sparse online learning via truncated gradient. In D. Koller, D. Schuurmans,

Y. Bengio, and L. Bottou (Eds.), *Advances in Neural Information Processing Systems*, Volume 21. Curran Associates, Inc.

Ledoit, O. and M. Wolf (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis 88*(2), 365–411.

Lee, J. D., Y. Sun, and J. E. Taylor (2015). On model selection consistency of regularized M-estimators. *Electronic Journal of Statistics 9*(1), 608 – 642.

Lee, K. E., N. Sha, E. R. Dougherty, M. Vannucci, and B. K. Mallick (2003, 01). Gene selection: a Bayesian variable selection approach. *Bioinformatics 19*(1), 90–97.

Li, Y., H. Yang, H. Yu, H. Huang, and Y. Shen (2023, 04). Penalized weighted least-squares estimate for variable selection on correlated multiply imputed data. *Journal of the Royal Statistical Society Series C: Applied Statistics*. qlad028.

Liang, H. and R. Li (2009). Variable selection for partially linear models with measurement errors. *Journal of the American Statistical Association 104*(485), 234–248. PMID: 20046976.

Liu, H., L. Wang, and T. Zhao (2014). Sparse covariance matrix estimation with eigenvalue constraints. *Journal of Computational and Graphical Statistics 23*(2), 439–459. PMID: 25620866.

Loh, P.-L. and M. J. Wainwright (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics 40*(3), 1637 – 1664.

Loh, P.-L. and M. J. Wainwright (2013). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger (Eds.), *Advances in Neural Information Processing Systems*, Volume 26. Curran Associates, Inc.

Loh, P.-L. and M. J. Wainwright (2017). Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics 45*(6), 2455 – 2482.

Long, Q. and B. A. Johnson (2015, 02). Variable selection in the presence of missing data: resampling and imputation.

Biostatistics 16(3), 596–610.

Lounici, K. (2014, 08). High-dimensional covariance matrix estimation with missing observations. *Bernoulli 20*(3), 1029–1058.

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

Mendelson, S. and N. Zhivotovskiy (2020). Robust covariance estimation under $L_4 - L_2$ norm equivalence. *The Annals of Statistics 48*(3), 1648 – 1664.

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.

Omenn, G. S., Y. Guan, and R. Menon (2014). A new class of protein cancer biomarker candidates: Differentially expressed splice variants of erbb2 (her2/neu) and erbb1 (egfr) in breast cancer cell lines. *Journal of Proteomics 107*, 103–112. Special Issue: "20 years of Proteomics" in memory of Vitaliano Pallini.

Osborne, M. R., B. Presnell, and B. A. Turlach (2000). On the lasso and its dual. *Journal of Computational and Graphical Statistics 9*(2), 319–337.

Park, S. and J. Lim (2019). Non-asymptotic rate for high-dimensional covariance estimation with non-independent missing observations. *Statistics & Probability Letters 153*, 113–123.

Park, S., X. Wang, and J. Lim (2021). Estimating high-dimensional covariance and precision matrices under general missing dependence. *Electronic Journal of Statistics 15*(2), 4868 – 4915.

Park, S., X. Wang, and J. Lim (2023). Sparse HansonWright inequality for a bilinear form of sub-gaussian variables. *Stat 12*(1), e539.

Pavez, E. and A. Ortega (2021). Covariance matrix estimation with non uniform and data dependent missing observations. *IEEE Transactions on Information Theory 67*(2), 1201–1215.

Qin, Y., A. P. Conley, E. A. Grimm, and J. Roszik (2017, 04). A tool for discovering drug sensitivity and gene

expression associations in cancer cells. *PLOS ONE 12*(4), 1–6.

Romeo, G. and M. Thoresen (2019). Model selection in high-dimensional noisy data: a simulation study. *Journal of Statistical Computation and Simulation 89*(11), 2031–2050.

Rosenbaum, M. and A. B. Tsybakov (2010). Sparse recovery under matrix uncertainty. *The Annals of Statistics 38*(5), 2620 – 2651.

Rothman, A. J. (2012, 06). Positive definite estimators of large covariance matrices. *Biometrika 99*(3), 733–740.

Sørensen, Ø. (2019). hdme: High-dimensional regression with measurement error. *Journal of Open Source Software 4*(37), 1404.

Sørensen, Ø., A. Frigessi, and M. Thoresen (2015). Measurement error in lasso: Impact and likelihood bias correction. *Statistica Sinica 25*, 809829.

Stdler, N. and P. Bhlmann (2010). Missing values: Sparse inverse covariance estimation and extension to sparse regression. *Statistics and Computing 22*(1), 219235.

Takada, M., H. Fujisawa, and T. Nishikawa (2019, 7). Hmlasso: Lasso with high missing rate. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 3541–3547. International Joint Conferences on Artificial Intelligence Organization.

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

van de Geer, S. A. and P. Bühlmann (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics 3*, 1360 – 1392.

Vershynin, R. (2011). Introduction to the non-asymptotic analysis of random matrices.

Vis, D. J., L. Bombardelli, H. Lightfoot, F. Iorio, M. J. Garnett, and L. F. Wessels (2016). Multilevel models improve precision and speed of ic50 estimates. *Pharmacogenomics 17*(7), 691–700. PMID: 27180993.

Wainwright, M. J. (2009). 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.

Wan, Y., S. Datta, D. Conklin, and M. Kong (2015). Variable selection models based on multiple imputation with an application for predicting median effective dose and maximum effect. *Journal of Statistical Computation and Simulation 85*(9), 1902–1916.

Wang, Y., J. Wang, S. Balakrishnan, and A. Singh (2019). Rate optimal estimation and confidence intervals for high-dimensional regression with missing covariates. *Journal of Multivariate Analysis 174*, 104526.

Webb-Robertson, B.-J. M., H. K. Wiberg, M. M. Matzke, J. N. Brown, J. Wang, J. E. McDermott, R. D. Smith, K. D. Rodland, T. O. Metz, J. G. Pounds, and K. M. Waters (2015). Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *Journal of Proteome Research 14*(5), 1993–2001. PMID: 25855118.

Wei, R., J. Wang, M. Su, E. Jia, S. Chen, T. Chen, and Y. Ni (2018, Jan). Missing value imputation approach for mass spectrometry-based metabolomics data. *Scientific Reports 8*(1), 663.

Wood, A. M., I. R. White, and P. Royston (2008). How should variable selection be performed with multiply imputed data? *Statistics in Medicine 27*(17), 3227–3246.

Wu, L., S. I. Candille, Y. Choi, D. Xie, L. Jiang, J. Li-Pook-Than, H. Tang, and M. Snyder (2013, Jul). Variation and genetic control of protein abundance in humans. *Nature 499*(7456), 79–82.

Xiao, L. (2009). Dual averaging method for regularized stochastic learning and online optimization. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta (Eds.), *Advances in Neural Information Processing Systems*, Volume 22. Curran Associates, Inc.

Xu, Z.-q., Y. Zhang, N. Li, P.-j. Liu, L. Gao, X. Gao, and X.-j. Tie (2017). Efficacy and safety of lapatinib and trastuzumab for her2-positive breast cancer: a systematic review and meta-analysis of randomised controlled trials. *BMJ Open 7*(3).

Xue, L., S. Ma, and H. Zou (2012). Positive-definite ℓ_1-penalized estimation of large covariance matrices. *Journal of the American Statistical Association 107*(500), 1480–1491.

Zhang, J., Y. Li, N. Zhao, and Z. Zheng (2022). L0-regularization for high-dimensional regression with corrupted data. *Communications in Statistics - Theory and Methods 0*(0), 1–17.

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

Zheng, L. and G. I. Allen (2023). Graphical model inference with erosely measured data. *Journal of the American Statistical Association 0*(ja), 1–22.

Zheng, Z., Y. Li, C. Yu, and G. Li (2018). Balanced estimation for high-dimensional measurement error models. *Computational Statistics & Data Analysis 126*, 78–91.

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

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

Seongoh Park, School of Mathematics, Statistics and Data Science, Sungshin Women's University

E-mail: spark6@sungshin.ac.kr

Seong Jin Lee, Department of Statistics and Operations Research, University of North Carolina at Chapel Hill

E-mail: slee7@unc.edu

Nguyen Thi Hai Yen, Department of Pharmacology and PharmacoGenomics Research Center, Inje University College of Medicine

E-mail: haiyenzenny97a1@gmail.com

Nguyen Phuoc Long, Graduate Institute of Biomedical Sciences, College of Medicine, Chang Gung University

E-mail: phuoclong@mail.cgu.edu.tw

Johan Lim, Department of Statistics, Seoul National University

E-mail: johanlim@snu.ac.kr