HMLasso: Lasso with High Missing Rate

Masaaki Takada[1*], **Hironori Fujisawa**[2] and **Takeichiro Nishikawa**[1]
[1]Toshiba Corporation
[2]The Institute of Statistical Mathematics
masaaki1.takada@toshiba.co.jp, fujisawa@ism.ac.jp, takeichiro.nishikawa@toshiba.co.jp

Abstract

Sparse regression such as the Lasso has achieved great success in handling high-dimensional data. However, one of the biggest practical problems is that high-dimensional data often contain large amounts of missing values. Convex Conditioned Lasso (CoCoLasso) has been proposed for dealing with high-dimensional data with missing values, but it performs poorly when there are many missing values, so that the high missing rate problem has not been resolved. In this paper, we propose a novel Lasso-type regression method for high-dimensional data with high missing rates. We effectively incorporate mean imputed covariance, overcoming its inherent estimation bias. The result is an optimally weighted modification of CoCoLasso according to missing ratios. We theoretically and experimentally show that our proposed method is highly effective even when there are many missing values.

1 Introduction

High-dimensional data appear in a wide range of fields, including biology, economy, and industry. Over the past several decades, sparse regression has achieved great success in dealing with high-dimensional data, because it efficiently performs both model estimation and variable selection simultaneously. Sparse regression methods include the Lasso [Tibshirani, 1996], Elastic Net [Zou and Hastie, 2005], SCAD [Fan and Li, 2001], and MCP [Zhang, 2010].

In practice, high-dimensional data often contain large amounts of missing values. For example, educational and psychological studies commonly have missing data ratios of 15–20% [Enders, 2003], while maintenance data for typical industrial processes had over 75% missing values in over 50% of variables [Lakshminarayan *et al.*, 1999]. Up to 90% of traffic data can be missing [Tan *et al.*, 2013]. There is thus demand for methods that can accommodate high-dimensional data with a high missing rate.

Missing data analysis has a long history. Listwise deletion (complete case analysis) and pairwise deletion are widely used because of their simplicity. Various methods have been developed, including the expectation maximization (EM) algorithm [Dempster *et al.*, 1977], multiple imputation (MI) [Rubin, 1987; Schafer, 1997; Buuren and Groothuis-Oudshoorn, 2011], and full information maximum likelihood (FIML) [Hartley and Hocking, 1971; Enders, 2001]. However, these methods focus on low-dimensional missing data and are intractable for high-dimensional data due to their computational cost.

To deal with high-dimensional missing data, a direct regression method using a pairwise covariance matrix has been proposed [Loh and Wainwright, 2012]. This method incurs low computational costs, but heavily depends on some critical unknown parameters that must be determined in advance due to its nonconvexity. Convex Conditioned Lasso (CoCoLasso) was proposed to avoid this problem [Datta and Zou, 2017]. Because of its convexity using a positive semidefinite approximation of the pairwise covariance matrix, it does not suffer from local optima or critical parameters. However, we found that CoCoLasso can deteriorate at high missing rates. Indeed, estimator accuracy may significantly worsen even when only one variable has a high missing rate, so a high missing rate remains problematic.

In this paper, we propose a novel regression method that overcomes problems related to both high-dimensionality and high missing rates. We use the mean imputed covariance matrix, effectively incorporating it into Lasso despite its noted tendency toward estimation bias for missing data. The resulting optimization problem can be seen as a weighted modification of CoCoLasso using the missing ratio, and it is quite effective for high-dimensional data with a high missing rate. Our proposed method is free from local optima and critical parameters due to convexity, and it is theoretically and experimentally superior to CoCoLasso regarding estimation error. Contributions of this study are as follows:

- We propose a novel regression method for handling high-dimensional data with high missing rates.
- We analyze theoretical properties of our method, showing that our formulation is superior to all other weighted formulations with regards to estimation error.
- We demonstrate the effectiveness of our method through both numerical simulations and real-world data experiments. Our method outperforms other methods in al-

*Contact Author

most all cases, and particularly shows significant improvement for high-dimensional data with high missing rates.

The remainder of this paper is organized as follows: We first review existing methods and describe our proposed method. We then show the advantages of our method through theoretical analyses, numerical simulations, and real-world data experiments.

1.1 Notations

Let $v \in \mathbb{R}^p$. $\|v\|_q$ $(q > 0)$ is the ℓ_q norm, that is, $\|v\|_q = (|v_1|^q + \cdots + |v_p|^q)^{1/q}$. Let $M \in \mathbb{R}^{n\times p}$. $\|M\|_{\mathrm{F}}$ is the Frobenius norm, that is, $\|M\|_{\mathrm{F}} = (\sum_{j,k} M_{jk}^2)^{1/2}$. $\|M\|_{\max}$ is the max norm, that is, $\|M\|_{\max} = \max_{j,k} |M_{jk}|$. Let $M_1, M_2 \in \mathbb{R}^{n\times p}$. $M_1 \odot M_2$ is the element-wise product (Hadamard product) of M_1 and M_2. $M_1 \oslash M_2$ is the element-wise division of M_1 and M_2. Let $M \in \mathbb{R}^{p\times p}$ be a symmetric matrix. $M \succeq 0$ denotes that M is positive semidefinite, that is, $v^\top M v \geq 0$ for any $v \in \mathbb{R}^p$.

2 Methods

2.1 Problem Formulation

Consider a linear regression model $y = X\beta + \varepsilon$, where $X \in \mathbb{R}^{n\times p}$ is a complete design matrix, $y \in \mathbb{R}^n$ is a response, $\beta \in \mathbb{R}^p$ is a regression coefficient, and $\varepsilon \in \mathbb{R}^n$ is a noise. Suppose that y and each column of X are centered without loss of generality. The ordinary problem is to estimate the regression coefficient β given complete data X and y. In contrast, we consider the situation where some elements of X are missing in this paper.

If X does not contain missing values, Lasso is one of the most promising methods for handling high-dimensional data. It solves the problem [Tibshirani, 1996]

$$\hat{\beta} = \operatorname*{argmin}_{\beta} \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1, \tag{1}$$

where $\lambda > 0$ is a regularization parameter. Since the objective function (1) is regularized by the ℓ_1 norm of β, the solution is sparse and often has a small generalization error. In the presence of missing values, however, it is impossible to directly apply the Lasso.

2.2 Review of Existing Methods

Interestingly, to estimate the parameter β in the presence of missing values does *not* require imputation in X, which is computationally expensive for high-dimensional data. We can directly estimate the parameter without imputation. The Lasso objective function (1) can be reformulated as

$$\hat{\beta} = \operatorname*{argmin}_{\beta} \frac{1}{2}\beta^\top S\beta - \rho^\top\beta + \lambda\|\beta\|_1, \tag{2}$$

where $S = \frac{1}{n}X^\top X$ (the sample covariance matrix of X) and $\rho = \frac{1}{n}X^\top y$ (the sample covariance vector of X and y). Using (2), we can estimate β via S and ρ instead of X and y. If data are missing completely at random, we can easily construct unbiased estimators of the covariance matrix and vector using the pairwise covariance, that is, $S^{\mathrm{pair}} = (S_{jk}^{\mathrm{pair}})$ and $\rho^{\mathrm{pair}} = (\rho_{jk}^{\mathrm{pair}})$ as

$$S_{jk}^{\mathrm{pair}} := \frac{1}{n_{jk}} \sum_{i\in I_{jk}} X_{ij}X_{ik}, \text{ and } \rho_j^{\mathrm{pair}} := \frac{1}{n_{jj}} \sum_{i\in I_{jj}} X_{ij}y_i,$$

where $I_{jk} := \{i : X_{ij} \text{ and } X_{ik} \text{ are observed }\}$, and n_{jk} is the number of elements of I_{jk}. Thus, we can replace S and ρ by S^{pair} and ρ^{pair} in (2), respectively.

The major problem here is that S^{pair} may *not* be positive semidefinite (PSD). In other words, it may have negative eigenvalues. This is a critical problem because negative eigenvalues can cause the objective function to diverge to minus infinity, meaning the optimization failed. A regression method with a nonconvex constrained formulation has been proposed to avoid this problem [Loh and Wainwright, 2012]. However, this method's nonconvexity causes some difficulties in practice. Namely, different initial values result in multiple global or local minimizers that produce different output solutions, and the solutions depend on unknown parameters that must be determined in advance. To avoid the above difficulties, a convex optimization problem called CoCoLasso has been proposed [Datta and Zou, 2017]. This problem is formulated as

$$\hat{\beta} = \operatorname*{argmin}_{\beta} \frac{1}{2}\beta^\top \check{\Sigma}\beta - {\rho^{\mathrm{pair}}}^\top\beta + \lambda\|\beta\|_1, \tag{3}$$

$$\check{\Sigma} = \operatorname*{argmin}_{\Sigma\succeq 0} \|\Sigma - S^{\mathrm{pair}}\|_{\max}. \tag{4}$$

CoCoLasso obtains the PSD covariance matrix in (4) via the alternating direction method of multipliers (ADMM) algorithm, then optimizes the Lasso objective function (3). This formulation overcomes the difficulties seen in [Loh and Wainwright, 2012] because the objective function (3) is convex due to the PSD matrix $\check{\Sigma}$, meaning it has no local minimizers, and because it uses no unknown parameters that must be determined in advance. In addition, statistical non-asymptotic properties were also derived. For these reasons, CoCoLasso is practical and state-of-the-art.

However, a high missing rate can deteriorate estimations of the covariance matrix in CoCoLasso. If some pairwise observation numbers n_{jk} are very small, then the corresponding pairwise covariances S_{jk}^{pair} are quite unreliable, possibly becoming very large or small. Since (4) is based on the max norm, unreliable elements of S^{pair} will greatly affect the estimator. As a result, other estimator elements can highly deviate from the corresponding elements in S^{pair}, even if their variables have few missing values. This indicates that CoCoLasso results can significantly worsen, even if only one variable has a high missing rate. The problem is that CoCoLasso does not account for the differences in reliability of the pairwise covariance. The next subsection describes how we overcome this problem.

We mention other approaches for regression with missing data. A simple approach is listwise deletion. This method is very fast but inappropriate when there are few complete samples, as is common with high-dimensional data. Another typical approach is to impute missing values, including the

mean imputation, the EM algorithm [Dempster *et al.*, 1977], MI [Rubin, 1987; Schafer, 1997; Buuren and Groothuis-Oudshoorn, 2011], FIML [Hartley and Hocking, 1971; Enders, 2001], and other non-parametric methods [Stekhoven and Bühlmann, 2012]. These methods, however, typically incur large computational costs or cause large bias in high-dimensional problems, and it is hard to conduct theoretical analysis with these methods. Other direct modeling methods use the Dantzig selector instead of the Lasso [Rosenbaum and Tsybakov, 2010; Rosenbaum and Tsybakov, 2013]. An advantage of the Lasso-type approach is that computation is empirically much faster than with the Dantzig selector [Efron *et al.*, 2007].

2.3 Proposed Method: HMLasso

The mean imputation method is commonly used in practice. Let Z be the mean imputed data of X. Because X is centered, $Z_{jk} = X_{jk}$ for observed elements and $Z_{jk} = 0$ for missing elements. The covariance matrix of the mean imputed data, $S^{\mathrm{imp}} = (S^{\mathrm{imp}}_{jk})$, is defined as

$$S^{\mathrm{imp}}_{jk} = \frac{1}{n}\sum_{i=1}^{n} Z_{ij}Z_{ik} = \frac{n_{jk}}{n}\frac{1}{n_{jk}}\sum_{i\in I_{jk}} X_{ij}X_{ik} = \frac{n_{jk}}{n}S^{\mathrm{pair}}_{jk}. \tag{5}$$

We can equivalently express (5) as

$$S^{\mathrm{imp}} = R \odot S^{\mathrm{pair}} \tag{6}$$

where $R = (R_{jk})$ with $R_{jk} = n_{jk}/n$. The mean imputed covariance matrix S^{imp} is biased but PSD, while the pairwise covariance matrix S^{pair} is unbiased but not PSD. To take the best aspects of both, we optimize

$$\tilde{\Sigma} = \underset{\Sigma\succeq 0}{\operatorname{argmin}} \|R \odot \Sigma - S^{\mathrm{imp}}\|_{\mathrm{F}}^2 \tag{7}$$

to obtain a low-biased and PSD matrix. Direct use of mean imputation for covariance matrix estimation is known to produce estimation bias. However, we can use the relation (6) between S^{imp} and S^{pair} to effectively incorporate them into the optimization problem (7).

The formulation (7) has a useful property, in that it is equivalent to

$$\tilde{\Sigma} = \underset{\Sigma\succeq 0}{\operatorname{argmin}} \|R \odot (\Sigma - S^{\mathrm{pair}})\|_{\mathrm{F}}^2. \tag{8}$$

The formulation (8) can be seen as a weighted modification of CoCoLasso (4) using the observed ratio matrix R, where the max norm is replaced by the Frobenius norm. This weighting is beneficial under high missing rates. When there are missing observations, the objective function downweights the corresponding term $\Sigma_{jk} - S^{\mathrm{pair}}_{jk}$ by the observed ratio $R_{jk} = n_{jk}/n$. In particular, the downweighting will be reasonable when n_{jk} is small, because the pairwise covariance S^{pair}_{jk} is unreliable.

From the above, we extend the formulation and propose a novel optimization problem to estimate the regression model as

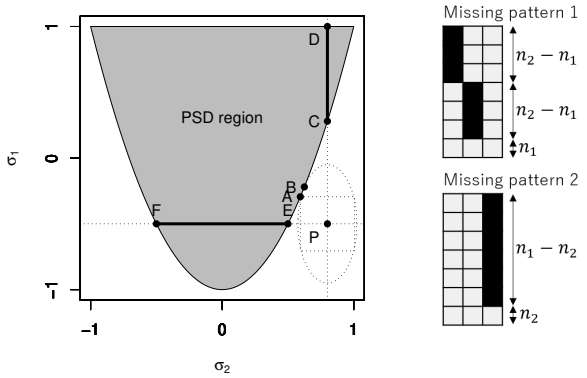

Figure 1: (Left) Two-dimensional covariance matrix space. (Right) Two simple missing patterns. Black elements represent missing values.

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2}\beta^\top\tilde{\Sigma}\beta - {\rho^{\mathrm{pair}}}^\top\beta + \lambda\|\beta\|_1, \tag{9}$$

$$\tilde{\Sigma} = \underset{\Sigma\succeq 0}{\operatorname{argmin}} \|W \odot (\Sigma - S^{\mathrm{pair}})\|_{\mathrm{F}}^2, \tag{10}$$

where W is a weight matrix whose (j,k)-th element is $W_{jk} = R^{\alpha}_{jk}$ with a constant $\alpha \geq 0$. We obtain a PSD matrix by minimizing the weighted Frobenius norm in (10), then optimizes the Lasso problem (9). This Lasso-type formulation allows us to efficiently deal with high missing rates, so we call our method "HMLasso".

Several α values can be considered. Setting $\alpha = 0$ corresponds to the non-weighted case, which is just a projection of S^{pair} onto the PSD region. This is the same as CoCoLasso when the Frobenius norm in (10) is replaced by the max norm. The case where $\alpha = 1$ relates to mean imputation, as described above. As shown below, non-asymptotic analyses support that $\alpha = 1$ is reasonable, and numerical experiments show that this setting delivers the best performance. Therefore, we recommend setting $\alpha = 1$ in practice. The case where $\alpha = 1/2$ can be roughly viewed as the maximum likelihood method from an asymptotic perspective. We discuss the case of setting $\alpha = 1/2$ in the supplementary material[1]. Note that in (10), we use the Frobenius norm instead of the max norm, because the Frobenius norm delivered better performance in numerical experiments.

2.4 Comparison Using a Simple Example

The following simple example shows that our weighted formulation (10) is better than the non-weighted formulation. Consider three-dimensional data $X \in \mathbb{R}^{n\times 3}$. To derive simple analytical results, we suppose that the pairwise covariance matrix and observation ratio matrix are

$$S^{\mathrm{pair}} = \begin{bmatrix} 1 & s_1 & s_2 \\ s_1 & 1 & s_2 \\ s_2 & s_2 & 1 \end{bmatrix} \text{ and } R = \frac{1}{n}\begin{bmatrix} \cdot & n_1 & n_2 \\ n_1 & \cdot & n_2 \\ n_2 & n_2 & \cdot \end{bmatrix},$$

[1]The whole paper including the supplementary material is available at https://arxiv.org/abs/1811.00255.

Algorithm 1 Covariance Estimation with ADMM

Input: $S^{\mathrm{pair}}, W, \mu$
  for $k = 1, 2, \ldots$ **do**
    $A_{k+1} \leftarrow$ projection of $B_k + S^{\mathrm{pair}} + \mu\Lambda_k$ onto PSD
    $B_{k+1} \leftarrow (A_{k+1} - S^{\mathrm{pair}} - \mu\Lambda_k) \oslash (\mu W \odot W + 11^\top)$
    $\Lambda_{k+1} \leftarrow \Lambda_k - \frac{1}{\mu}\left(A_{k+1} - B_{k+1} - S^{\mathrm{pair}}\right)$
  end for
Output: A_{i+1}

respectively. We restrict diagonal elements of the covariance estimate to 1 for simplicity. From symmetry of the problem, we can parameterize the covariance estimate as $\Sigma = [1, \sigma_1, \sigma_2; \sigma_1, 1, \sigma_2; \sigma_2, \sigma_2, 1]$. Here, we can see S^{pair} and Σ in the same two-dimensional space as in Figure 1. Simple calculation yields that the PSD condition of S^{pair} is $2s_2^2 - 1 \leq s_1 \leq 1$, shown in gray in Figure 1. Hereafter, to show differences among the methods, without loss of generality we suppose $s_1 < 2s_2^2 - 1$ so that S^{pair} is not PSD, and also $s_2 \geq 0$. Point P in Figure 1 is an example of such an S^{pair}.

Consider the case where n_1 is sufficiently small and n_2 is sufficiently large, which is realized by missing data pattern 1 in Figure 1. The CoCoLasso (non-weighted max norm) solution is the tangent point of the gray region and an elliptic contour with center point P, shown as point A in Figure 1. The solution of the non-weighted Frobenius norm is the tangent point of the gray region and a square contour with center point P, shown as point B. The optimal point of HMLasso (weighted norm) will be close to point C, because weights R_{13} and R_{23} are much larger than R_{12}, so the optimal solution must be close to the tangent point of the gray PSD region and line CD. On the other hand, the sample covariance matrix S with complete observations satisfies $\sigma_2 \approx s_2$ and $2\sigma_2^2 - 1 \leq \sigma_1 \leq 1$, represented as line segment CD. Since point C is closer to any point on line segment CD than are A or B, the HMLasso estimate is always closer to the covariance matrix of the complete data than are estimates using non-weighted norms.

Consider another case where n_1 is sufficiently large and n_2 is sufficiently small, which is realized by missing pattern 2 in Figure 1. By reasoning similar to the case described above, the solutions for the non-weighted max and Frobenius norm are A and B, respectively, and the HMLasso optimal point will be close to point E in Figure 1. The sample covariance matrix S with complete observations satisfies $\sigma_1 \approx s_1$ and $2\sigma_2^2 - 1 \leq \sigma_1 \leq 1$, shown as line segment EF. Hence, the HMLasso estimate is again superior to those of the non-weighted norms.

2.5 Algorithms

The Lasso optimization problem using the covariance matrix (9) can be solved by various algorithms for the Lasso, such as the coordinate descent algorithm [Friedman *et al.*, 2010] and the least angle regression algorithm [Efron *et al.*, 2004]. Our implementation uses the coordinate descent algorithm because it is efficient for high-dimensional data. The algorithm details are described in the supplementary material. We use the warm-start and safe screening techniques to speed up the algorithm.

We use the ADMM algorithm [Boyd *et al.*, 2011] to derive the PSD covariance matrix optimization (10), which can be rewritten as

$$(A, B) = \underset{A \succeq 0, B = A - S^{\mathrm{pair}}}{\operatorname{argmin}} \|W \odot B\|_{\mathrm{F}}^2.$$

Therefore, the augmented Lagrangian function is

$$f(A, B, \Lambda) = \frac{1}{2}\|W \odot B\|_{\mathrm{F}}^2 - \langle \Lambda, A - B - S^{\mathrm{pair}} \rangle + \frac{1}{2\mu}\|A - B - S^{\mathrm{pair}}\|_{\mathrm{F}}^2, \quad (11)$$

where Λ is a Lagrangian matrix and μ is an augmented Lagrangian parameter. We iteratively update A, B, and Λ subject to $A \succeq 0$ by minimizing (11) in terms of each variable. The resulting algorithm is similar to the CoCoLasso algorithm, except for the update rule for B due to weight matrix W. To derive the B-step update equation, differentiating $f(A, B, \Lambda)$ with respect to B yields

$$\partial_B f(A, B, \Lambda) = W \odot W \odot B + \Lambda - \frac{1}{\mu}\left(A - B - S^{\mathrm{pair}}\right).$$

Solving $\partial_B f(A, B, \Lambda) = 0$, we obtain the update rule

$$B \leftarrow (A - S^{\mathrm{pair}} - \mu\Lambda) \oslash (\mu W \odot W + 11^\top),$$

where $11^\top$ is a matrix of ones. The algorithm for solving (10) is presented as Algorithm 1. The difference between the max norm and the Frobenius norm is trivial when we use ADMM. The supplementary material also describes the algorithm for the weighted max norm.

3 Theoretical Properties

In this section, we investigate statistical properties of the proposed estimator. We first obtain a refined non-asymptotic property for the pairwise covariance matrix, which explicitly includes missing rate effects. We then derive a non-asymptotic property of our estimate in (10). These results show that $\alpha = 1$ weighting is superior in terms of non-asymptotic properties over other weighting ($\alpha \neq 1$) including non-weighted formulation ($\alpha = 0$). Note that we focus on the Frobenius norm formulation in (10), but we can see that $\alpha = 1$ weighting is superior as well for the max norm formulation, though CoCoLasso uses the non-weighted norm ($\alpha = 0$). Complete proofs for propositions and theorems in this section are given in the supplementary material.

3.1 Preliminaries

Let $M = (M_{ij}) \in \mathbb{R}^{n \times p}$ be the observation pattern matrix whose elements are 1 when data are observed and 0 otherwise, so that $Z = M \odot X$. We suppose a sub-Gaussian assumption on M, which plays a key role in non-asymptotic properties. This assumption often holds, as seen in the following proposition.

Definition 1. *A random variable $u \in \mathbb{R}$ is said to be sub-Gaussian with τ^2 if $E\left[\exp(s(u - E[u]))\right] \leq \exp\left(\tau^2 s^2/2\right)$ for all $s \in \mathbb{R}$. A random vector $u \in \mathbb{R}^p$ is said to be sub-Gaussian with τ^2 if $v^\top u$ is a sub-Gaussian variable with τ^2 for all $v \in \mathbb{R}^p$ satisfying $\|v\|_2 = 1$.*

Assumption 1. *The rows of M are independent and identically distributed with mean μ_M, covariance Σ_M, and sub-Gaussian parameter τ^2.*

Proposition 1. *Assume that M_{ij} values are independent and identically distributed as a Bernoulli distribution with mean μ_j. Then, the rows of M are sub-Gaussian with $\tau^2 = \max_j \mu_j(1-\mu_j) \leq 1/4$.*

For the theoretical analysis in this section, we substitute $\hat{\Sigma} := \frac{1}{n} Z^\top Z \oslash \Pi$ for S^{pair}, where $\Pi := (\pi_{jk}) := \Sigma_M + \mu_M \mu_M^\top$. This is reasonable because $S^{\text{pair}} = \frac{1}{n} Z^\top Z \oslash R$ from (5), and the expectation for R is $\mathrm{E}[R] = \mathrm{E}[\frac{1}{n} M^\top M] = \Pi$. We respectively call π_{jk} and $1 - \pi_{jk}$ the observed and missing rate, since they are expectations for the observed and missing ratios. We suppose that μ_M and Σ_M are known. This substitution and assumption were also used in previous theoretical research [Loh and Wainwright, 2012; Datta and Zou, 2017].

3.2 Statistical Properties

We first derive a refined non-asymptotic property of $\hat{\Sigma}$.

Theorem 2. *Under Assumption 1, we have, for all $\varepsilon \leq c\tau^2 X_{\max}^2 / \pi_{jk}$,*

$$\Pr\left(\left|\hat{\Sigma}_{jk} - S_{jk}\right| \leq \varepsilon\right) \geq 1 - C\exp\left(-cn\varepsilon^2 \pi_{jk}^2 \zeta^{-1}\right),$$

where $\zeta = \tau^2 X_{\max}^4 \max\left\{\tau^2, \mu_j^2, \mu_k^2\right\}$, $X_{\max} = \max_{i,j} |X_{ij}|$, and C and c are universal constants.

Sketch of Proof. We see that

$$\left|\hat{\Sigma}_{jk} - S_{jk}\right| \leq \frac{1}{n\pi_{jk}} \left|\sum_{i=1}^n v_{ijk}(m_{ij} - \mu_j)(m_{ik} - \mu_k)\right| + \frac{\mu_j}{n\pi_{jk}} \left|\sum_{i=1}^n v_{ijk}(m_{ik} - \mu_k)\right| + \frac{\mu_k}{n\pi_{jk}} \left|\sum_{i=1}^n v_{ijk}(m_{ij} - \mu_j)\right|,$$

with $v_{ijk} := x_{ij}x_{ik}$. The first term is bounded using Lemma B.1 in [Datta and Zou, 2017], and the second and third terms are bounded using Property (B.2) in [Datta and Zou, 2017]. Careful analyses considering π_{jk}, μ_j, and μ_k yield the assertion. □

In Theorem 2, the missing rate appears explicitly and the non-asymptotic property is stricter than Definition 1 and Lemma 2 in [Datta and Zou, 2017]. To clearly see the missing rate effect, we replace ε by ε/π_{jk}. Then, for all $\varepsilon \leq c\tau^2 X_{\max}^2$ we have

$$\Pr\left(\pi_{jk}\left|\hat{\Sigma}_{jk} - S_{jk}\right| \leq \varepsilon\right) \geq 1 - C\exp\left(-cn\varepsilon^2\zeta^{-1}\right).$$

Since the right side does not depend on π_{jk}, we can see that the concentration probability of $\pi_{jk}|\hat{\Sigma}_{jk} - S_{jk}|$ is equally bounded regardless of the missing rate. This implies that our weighted formulation balances uncertainty of each element of $\hat{\Sigma}$, while non-weighted formulations such as CoCoLasso suffer from this imbalance.

Next, we derive a non-asymptotic property of our weighted estimator $\tilde{\Sigma}$.

Theorem 3. *Under Assumption 1, we have, for all $\varepsilon \leq c\tau^2 X_{\max}^2 (\min_{j,k} W_{jk}/\pi_{jk}) / W_{\min}$,*

$$\Pr\left(\frac{1}{p^2}\left\|\tilde{\Sigma} - S\right\|_{\mathrm{F}}^2 \leq \varepsilon^2\right) \geq 1 - p^2 C\exp\left(-cn\varepsilon^2 W_{\min}^2 \left(\min_{j,k}\frac{\pi_{jk}}{W_{jk}}\right)^2 \zeta^{-1}\right),$$

where $\zeta = \tau^2 X_{\max}^4 \max\left\{\tau^2, \mu_1^2, \ldots, \mu_p^2\right\}$, $W_{\min} = \min_{j,k} W_{jk}$, $X_{\max} = \max_{i,j}|X_{ij}|$, and C and c are universal constants.

Sketch of Proof. We have $\|W \odot (\tilde{\Sigma} - S)\|_{\mathrm{F}} \leq \|W \odot (\tilde{\Sigma} - \hat{\Sigma})\|_{\mathrm{F}} + \|W \odot (\hat{\Sigma} - S)\|_{\mathrm{F}} \leq 2\|W \odot (\hat{\Sigma} - S)\|_{\mathrm{F}}$ by the triangular equation and the optimality of $\tilde{\Sigma}$. Using $W_{\min}^2 \|\tilde{\Sigma} - S\|_{\mathrm{F}}^2 \leq \|W \odot (\tilde{\Sigma} - S)\|_{\mathrm{F}}^2$, Theorem 2 yields the assertion. □

According to Theorem 3, we can see that the proposed weight $W_{jk} = n_{jk}$ is optimal in the population level as follows.

Theorem 4. *Let $W_{jk} = \pi_{jk}^\alpha$. Then, the lower bound of the concentration probability and the upper bound of ε in Theorem 3 are minimized and maximized simultaneously when $\alpha = 1$.*

Proof. Let $\pi_{\min} = \min_{j,k} \pi_{jk}$ and $\pi_{\max} = \max_{j,k} \pi_{jk}$. Substituting $W_{jk} = \pi_{jk}^\alpha$, we have the concentration probability

$$1 - p^2 C\exp\left(-cn\varepsilon^2 \pi_{\min}^{2\alpha}\left(\min_{j,k}\pi_{jk}^{2-2\alpha}\right)\zeta^{-1}\right),$$

with the constraint $\varepsilon \leq c\tau^2 X_{\max}^2 (\min_{j,k}\pi_{jk}^{\alpha-1})/\pi_{\min}^\alpha$. i) For $0 \leq \alpha \leq 1$, the concentration probability becomes $1 - p^2 C\exp\left(-cn\varepsilon^2\pi_{\min}^2\zeta^{-1}\right)$, and the constraint of ε becomes $\varepsilon \leq c\tau^2 X_{\max}^2 (\pi_{\max}/\pi_{\min})^\alpha / \pi_{\max}$. Since $\pi_{\max}/\pi_{\min} \geq 1$, the constraint region of ε is maximized at $\alpha = 1$. ii) For $\alpha \geq 1$, the concentration probability becomes $1 - p^2 C\exp\left(-cn\varepsilon^2\pi_{\max}^2(\pi_{\min}/\pi_{\max})^{2\alpha}\right)$ and the constraint of ε becomes $\varepsilon \leq c\tau^2 X_{\max}^2/\pi_{\min}$. Since $\pi_{\min}/\pi_{\max} \leq 1$, the concentration probability is maximized at $\alpha = 1$. □

4 Numerical Experiments

We conducted some experiments using both synthetic and real-world data. More comprehensive simulation results under various conditions were given in the supplementary material due to space limitations.

4.1 Simulations with Various Norms

First, we investigated the effect of various weighted norms in (10). We compared the max norm with $\alpha = 0, 1/2, 1, 2$ and the Frobenius norm with $\alpha = 0, 1/2, 1, 2$. The max norm with $\alpha = 0$ corresponds to CoCoLasso, and the Frobenius norm with $\alpha = 0$ corresponds to a simple projection onto the PSD region.

Training data were $X \in \mathbb{R}^{n \times p}$ with $n = 10000$ and $p = 100$ generated by $\mathcal{N}(0, \Sigma^*)$ with $\Sigma_{jk}^* = 0.5$ for $j \neq k$ and $\Sigma_{jk}^* = 1$ for $j = k$. Responses y were defined as $y = X\beta + \varepsilon$

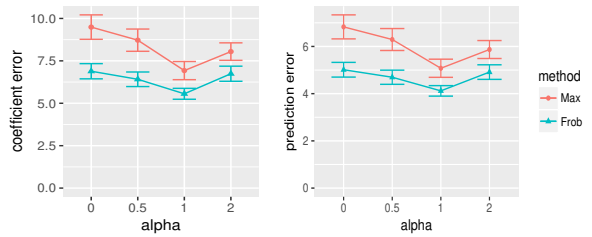

Figure 2: Comparison of the max and Frobenius norms with $\alpha = 0, 0.5, 1, 2$.

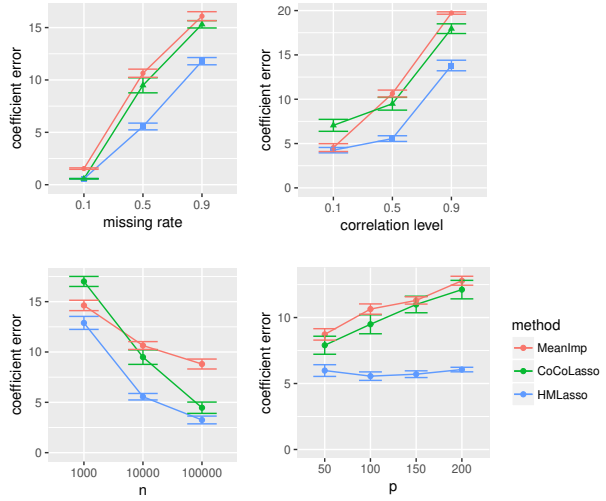

Figure 3: Simulation results with various missing rates, covariance levels, sample sizes, and dimension numbers.

with $\beta_1 = 10, \beta_{11} = -9, \beta_{21} = 8, \beta_{31} = -7, \ldots, \beta_{91} = -1$, and $\beta_j = 0$ otherwise, and $\varepsilon \sim \mathcal{N}(0, 1)$. We introduced missing values completely at random, setting a missing rate for each column sampled from a uniform distribution $U(0, 1)$ (0.5 on average). Test data were generated independently in the same manner, except that we did not introduce missing values for evaluation. The regularization parameter λ was selected by five-fold *corrected cross-validation* as with [Datta and Zou, 2017]. We iterated each experiment 30 times and plotted averaged results with standard errors.

Figure 2 shows the results. The performance measures were the ℓ_2 error for the regression coefficients, and the root mean square error of prediction. The weighted norms with $\alpha = 1$ were effective for both the Frobenius and max norm formulations, as suggested by the non-asymptotic theoretical analyses. In addition, the Frobenius norms outperformed the max norms. Therefore, we use the Frobenius norm with $\alpha = 1$ as the proposed method, namely, HMLasso, in subsequent experiments. In contrast, CoCoLasso (the max norm with $\alpha = 0$) was apparently inferior to HMLasso.

4.2 Simulations Under Various Conditions

We next compared the performance of HMLasso with other methods under various conditions, examining the mean imputation method, CoCoLasso (the max norm with $\alpha = 0$), and HMLasso (the Frobenius norm with $\alpha = 1$). Following the simulation setting in the previous subsection, we varied the missing data rate, covariance level, sample size, and number of variables. We set the average missing rates to $\mu = 0.1, 0.5, 0.9$, the covariance levels to $r = 0.1, 0.5, 0.9$, the sample size to $n = 10^3, 10^4, 10^5$, and the number of variables to $p = 50, 100, 150, 200$. Note that we also examined other missing imputation methods such as `mice` [Buuren and Groothuis-Oudshoorn, 2011] and `missForest` [Stekhoven and Bühlmann, 2012], but their computational costs were over 100 times larger than those for the above methods, so we excluded these methods in our experiments.

Figure 3 shows the results. HMLasso outperformed other methods under almost all conditions, especially for data with high missing rates and high covariance levels. In addition, high dimensionality did not adversely affect HMLasso, while the other methods showed gradually worse performance.

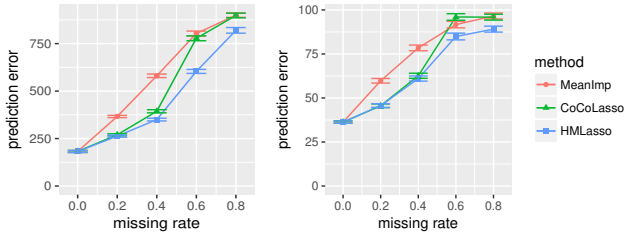

Figure 4: Analysis of residential building data with various patterns and rates for missing data. The outcomes are sales prices (left) and construction costs (right).

4.3 Residential Building Dataset

We evaluated the performance using a real-world residential building dataset [Rafiei and Adeli, 2016] from the UCI datasets repository[2]. The dataset included construction costs, sale prices, project variables, and economic variables corresponding to single-family residential apartments in Tehran. The objective was to predict sale prices and construction costs from physical, financial, and economic variables. The data consisted of $n = 372$ samples and $p = 105$ variables, including two output variables. We introduced missing values at missing rates $\mu = 0, 0.2, 0.4, 0.6, 0.8$ on average. We evaluated performance in terms of prediction error from complete samples. We randomly split data into 300 samples for training, 36 for validation, and 36 for testing, and iterated the experiments 30 times.

Figure 4 shows the results. HMLasso outperformed the other methods under nearly all cases. The advantage of HMLasso was clear especially for high missing rates.

5 Conclusion

We proposed a novel regression method for high-dimensional data with high missing rates, and demonstrated the advantages of our method through theoretical analyses and numerical experiments.

[2]https://archive.ics.uci.edu/ml/datasets/Residential+Building+Data+Set

References

[Boyd *et al.*, 2011] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[Buuren and Groothuis-Oudshoorn, 2011] S van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3), 12 2011.

[Datta and Zou, 2017] Abhirup Datta and Hui Zou. Cocolasso for high-dimensional error-in-variables regression. *The Annals of Statistics*, 45(6):2400–2426, 2017.

[Dempster *et al.*, 1977] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[Efron *et al.*, 2004] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[Efron *et al.*, 2007] Bradley Efron, Trevor Hastie, and Robert Tibshirani. Discussion: The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2358–2364, 2007.

[Enders, 2001] Craig K Enders. A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling*, 8(1):128–141, 2001.

[Enders, 2003] Craig K Enders. Using the expectation maximization algorithm to estimate coefficient alpha for scales with item-level missing data. *Psychological methods*, 8(3):322, 2003.

[Fan and Li, 2001] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[Friedman *et al.*, 2010] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[Hartley and Hocking, 1971] H O Hartley and R R Hocking. The analysis of incomplete data. *Biometrics*, pages 783–823, 1971.

[Lakshminarayan *et al.*, 1999] Kamakshi Lakshminarayan, Steven A Harp, and Tariq Samad. Imputation of missing data in industrial databases. *Applied intelligence*, 11(3):259–275, 1999.

[Loh and Wainwright, 2012] Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of statistics*, 40(3):1637–1664, 2012.

[Rafiei and Adeli, 2016] Mohammad Hossein Rafiei and Hojjat Adeli. A novel machine learning model for estimation of sale prices of real estate units. *Journal of Construction Engineering and Management*, 142(2):04015066, 2016.

[Rosenbaum and Tsybakov, 2010] Mathieu Rosenbaum and Alexandre B Tsybakov. Sparse recovery under matrix uncertainty. *The Annals of Statistics*, 38(5):2620–2651, 2010.

[Rosenbaum and Tsybakov, 2013] Mathieu Rosenbaum and Alexandre B Tsybakov. Improved matrix uncertainty selector. pages 276–290, 2013.

[Rubin, 1987] Donald B Rubin. *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, 1987.

[Schafer, 1997] Joseph L Schafer. *Analysis of incomplete multivariate data*. Chapman and Hall/CRC, 1997.

[Stekhoven and Bühlmann, 2012] Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

[Tan *et al.*, 2013] Huachun Tan, Guangdong Feng, Jianshuai Feng, Wuhong Wang, Yu-Jin Zhang, and Feng Li. A tensor-based method for missing traffic data completion. *Transportation Research Part C: Emerging Technologies*, 28:15–27, 2013.

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

[Zhang, 2010] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.

[Zou and Hastie, 2005] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.