Robust Sparse Estimation of Multiresponse Regression and Inverse Covariance Matrix via the L2 distance

Aurélie C. Lozano
IBM Watson Research Center
Yorktown Heights, New York
aclozano@us.ibm.com

Huijing Jiang
IBM Watson Research Center
Yorktown Heights, New York
huijiang@us.ibm.com

Xinwei Deng
Virginia Tech
Blacksburg, Virginia
xdeng@vt.edu

ABSTRACT

We propose a robust framework to jointly perform two key modeling tasks involving high dimensional data: (i) learning a sparse functional mapping from multiple predictors to multiple responses while taking advantage of the coupling among responses, and (ii) estimating the conditional dependency structure among responses while adjusting for their predictors. The traditional likelihood-based estimators lack resilience with respect to outliers and model misspecification. This issue is exacerbated when dealing with high dimensional noisy data. In this work, we propose instead to minimize a regularized distance criterion, which is motivated by the minimum distance functionals used in nonparametric methods for their excellent robustness properties. The proposed estimates can be obtained efficiently by leveraging a sequential quadratic programming algorithm. We provide theoretical justification such as estimation consistency for the proposed estimator. Additionally, we shed light on the robustness of our estimator through its linearization, which yields a combination of weighted lasso and graphical lasso with the sample weights providing an intuitive explanation of the robustness. We demonstrate the merits of our framework through simulation study and the analysis of real financial and genetics data.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning

Keywords

Robust estimation, high dimensional data, sparse learning, variable selection, multiresponse regression, inverse covariance, L2E

1. INTRODUCTION

We focus on multiresponse regression where both predictor and response spaces may exhibit high dimensions. We propose a *robust* framework to jointly and synergistically solve two important tasks: (i) learning the sparse functional mapping between inputs and outputs while taking advantage of the coupling among responses, and (ii) estimating the conditional dependency structure among responses while adjusting for the covariates. This is motivated by the crucial need of integrating genomic and transcriptomic datasets in computational biology in order to solve two fundamental problems effectively: identifying the genetic variations in the genome that influence gene expression levels (a.k.a. expression quantitative trait loci eQTLs mapping), and uncovering gene expression networks. In fact, the accuracy of the first problem can then be improved by leveraging the gene relatedness, and similarly the accurate and faithful estimation of the gene expression networks can be obtained by accounting for the confounding genetic effects on gene expression.

Multiresponse regression [5] generalizes the basic single-response regression to model multiple responses that might significantly correlate with each other. As opposed to treating each response independently, one can jointly learn multiple regression mappings to improve the estimation and prediction accuracy by exploiting the conditional dependencies among responses. Variable selection in multiresponse regression can be accomplished via the penalized approaches including lasso [31] and multitask lasso [24].

Sparse estimation of inverse covariance matrix is an important area in the multivariate analysis with broad applications in graphical models. A major focus in this area is that of penalized maximum likelihood formulations [15, 34, 1, 16]. Alternatively, modified Cholesky decompositions based on the likelihood can be used to estimate the sparse inverse covariance [17, 4, 21]. A simpler approach of "neighborhood selection" [22] estimates sparse graphical models using lasso to regress on each variable with the others as predictors.

Combining multiresponse regression and inverse covariance estimation has recently begun to attract more attention in the machine learning community. Rothman et al. [27] proposed a multivariate regression with covariance estimation (MRCE) to jointly estimate the sparse regression and inverse covariance matrices. They demonstrated that exploiting the correlation structures can significantly improve the prediction accuracy. The same model was also studied by Lee and Liu [20] who provided some theoretical properties for their developed method. An alternative parameterization was considered by Sohn and Kim [29], which is based on the joint distribution of predictors and responses and yields an l_1-penalized conditional graphical model (l_1-CGGM). Another relevant method is that of covariate adjusted precision matrix estimation (CAPME) [7], a two-stage approach

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.

to estimate the conditional dependency structure among response variables by adjusting for covariates. The first stage is to estimate the regression coefficients via a multivariate extension of the l_1 Dantzig selector [8], and the second stage is to estimate the inverse covariance matrix using l_∞ error with an l_1 penalty.

Robustness is an important aspect often overlooked in the sparse learning literature, while critical when dealing with high dimensional noisy data. Traditional likelihood-based estimators such as MRCE and l_1-CGGM lack resilience to outliers and model misspecification. Additionally, to the best of our knowledge, estimates based on Dantzig selector have not been compared to lasso counterparts in terms of robustness. Thus it is unclear whether CAPME, for instance, can address the robustness issue. There is limited existing work on robust sparse learning methods in high-dimensional modeling. The LAD-lasso [32] performs single response regression using the least absolute deviation combined with an l_1 penalty. The tlasso [14] performs inverse covariance estimation using penalized log-likelihood with the multivariate t distribution. However, neither of these methods can be easily extended to the setting of this paper.

We propose a robust approach to jointly estimate multiresponse regression and inverse covariance matrix. Our approach is based on a regularized distance criterion motivated by minimum distance estimators. Minimum distance estimators [33] are popularized in nonparametric methods and have exhibited excellent robustness properties [3, 12]. Their use for parametric estimation has been discussed in [28, 2]. In this work, we propose a penalized minimum distance criterion for robust estimation of sparse parametric models in the high dimensional settings. *Our key contributions* to this robust approach are as follows.

- The objective, which is denoted as REG-ISE, is based on the integrated squared error distance (ISE) between the model and the "true" distribution, and imposes the sparse model structure by adding sparsity-inducing penalties to the ISE criterion.
- Theoretical guarantees are provided on the estimation consistency of the proposed REG-ISE estimator.
- We leverage a sequential quadratic programming algorithm [9] to efficiently solve our objective.
- We shed light into the robustness of our framework by linearizing our objective. The linearization yields a problem combining weighted versions of l_1-penalized regression (lasso) and l_1-penalized inverse covariance estimation (glasso), where the weights assigned to the instances are theoretically derived and can be interpreted in terms of "outlying degrees".
- We propose a modified cross-validation and hold-out validation methods for the choice of tuning parameters, which are also applicable to other penalized regression methods.

The strength of our method is demonstrated via simulation data with and without outliers. Our study also confirms that outliers can severely influence the variable selection accuracy of some existing sparse learning methods. Experiments on real financial and eQTL data further illustrate the merits of the proposed method.

2. MODEL SETUP

Denote the response vector $\boldsymbol{y} = (y_1, \ldots, y_q)' \in \mathcal{R}^q$ and the predictor vector $\boldsymbol{x} = (x_1, \ldots, x_p)' \in \mathcal{R}^p$. We consider a multiresponse linear regression model

$$\boldsymbol{y} = \boldsymbol{B}'\boldsymbol{x} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{\Sigma}), \tag{1}$$

where $\boldsymbol{B} = (b_{ij})$ is a $p \times q$ matrix of coefficients and the kth column is the coefficients associated with kth response y_k regressing on the predictors $\boldsymbol{x}$. The $q \times q$ covariance matrix $\boldsymbol{\Sigma}$ describes the covariance structure of response vector $\boldsymbol{y}$ given the predictors $\boldsymbol{x}$. Moreover, its inverse $\boldsymbol{\Sigma}^{-1} = (c_{ij})$ represents the partial covariance structure [19] and has been widely used to learn a sparse graphical model under Gaussian assumption. Note that $\boldsymbol{\Sigma}^{-1}$ in (1) captures the partial covariances among responses $\boldsymbol{y}$ after adjusting for the effects of covariates $\boldsymbol{x}$. For simplicity of notation, we assume the data are centered so that the model (1) does not contain intercepts.

Suppose there are n observational vectors $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$, $i = 1, \ldots, n$ and the corresponding response vectors are $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iq})'$. To jointly obtain sparse estimates of coefficient matrix $\boldsymbol{B}$ and precision matrix $\boldsymbol{\Sigma}^{-1}$, we consider a loss function $L_n(\boldsymbol{B}, \boldsymbol{\Sigma})$ that measures the goodness-of-fit on the multivariate response. The sparse structures of $\boldsymbol{B}$ and $\boldsymbol{\Sigma}^{-1}$ are encouraged by using l_1 penalties. Specifically, the penalized loss function $L_{n,\lambda}(\boldsymbol{B}, \boldsymbol{\Sigma})$ is written as

$$L_{n,\lambda}(\boldsymbol{B}, \boldsymbol{\Sigma}) = L_n(\boldsymbol{B}, \boldsymbol{\Sigma}) + \lambda_1 \|\boldsymbol{\Sigma}^{-1}\|_1 + \lambda_2 \|\boldsymbol{B}\|_1. \tag{2}$$

where $\|\boldsymbol{\Sigma}^{-1}\|_1 = \sum_{i \le j} |c_{ij}|$ and $\|\boldsymbol{B}\|_1 = \sum_{i,j} |b_{ij}|$ are l_1 matrix norms. Following the principle of parsimony, we consider l_1 penalty functions to seek a most appropriate model that adequately explains the data. With carefully selected tuning parameters λ_1 and λ_2, we can achieve an optimal trade-off between the parsimoniousness and goodness-of-fit of the model.

The loss $L_n(\boldsymbol{B}, \boldsymbol{\Sigma})$ is typically derived from a likelihood-based approach. For instance the MRCE method [27] uses $L_n(\boldsymbol{B}, \boldsymbol{\Sigma}) = \text{trace}\left((\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})\boldsymbol{\Sigma}^{-1}\right) - \log \boldsymbol{\Sigma}^{-1}$. Altenatively, if one ignores the contribution of the inverse covariance matrix (by implicitely assuming that it is the identity matrix), one can consider the traditional squared loss $L_n = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_F^2$ and a straighforward generalization of the traditional Lasso estimator [31] to the mutiresponse setting.

3. A REGULARIZED INTEGRATED SQUARED ERROR ESTIMATOR

We begin this section by showing how a minimum distance criterion yields our proposed estimator for achieving robustness under the model in (1).

3.1 Derivation of the REG-ISE Objective

We first apply the Integrated Squared Error (ISE) criterion to the conditional distribution of response vector $\boldsymbol{y}$ given the predictors $\boldsymbol{x}$. It leads to an L_2 distance between the true conditional distribution $f(\boldsymbol{y}|\boldsymbol{x})$ and the parametric

distribution function $f(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{B},\boldsymbol{\Sigma})$ as follows

$$\begin{aligned}\tilde{L}(\boldsymbol{B},\boldsymbol{\Sigma}) &= \int \left[f(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{B},\boldsymbol{\Sigma}) - f(\boldsymbol{y}|\boldsymbol{x})\right]^2 d\boldsymbol{y} \\ &= \int f^2(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{B},\boldsymbol{\Sigma})d\boldsymbol{y} - 2\int f(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{B},\boldsymbol{\Sigma})f(\boldsymbol{y}|\boldsymbol{x})d\boldsymbol{y} \\ &+ \int f^2(\boldsymbol{y}|\boldsymbol{x})d\boldsymbol{y} \\ &= \int f^2(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{B},\boldsymbol{\Sigma})d\boldsymbol{y} - 2\mathbb{E}[f(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{B},\boldsymbol{\Sigma})] + \text{constant}.\end{aligned}$$

where $f(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{B},\boldsymbol{\Sigma})$ is the probability density function of multivariate normal $\mathcal{N}(\boldsymbol{B}'\boldsymbol{x},\boldsymbol{\Sigma})$ and $\int f(\boldsymbol{y}|\boldsymbol{x})^2 d\boldsymbol{y}$ is a constant independent of $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$.

Note that $f(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{B},\boldsymbol{\Sigma}) \equiv f(\boldsymbol{y}-\boldsymbol{B}'\boldsymbol{x};\boldsymbol{\Sigma})$ because of the conditional distribution assumption. Since $\boldsymbol{\epsilon} = \boldsymbol{y}-\boldsymbol{B}'\boldsymbol{x}$ are independently and identically distributed, one can consider approximating $\mathbb{E}[f(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{B},\boldsymbol{\Sigma})]$ by the empirical mean

$$\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{y}_i|\boldsymbol{x}_i;\boldsymbol{B},\boldsymbol{\Sigma}).$$

Similar approximation techniques have also been used for Gaussian mixture density estimation [28]. Now the resulting empirical loss function of $\tilde{L}(\boldsymbol{B},\boldsymbol{\Sigma})$ can therefore be written as

$$\tilde{L}_n(\boldsymbol{B},\boldsymbol{\Sigma}) = \int f^2(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{B},\boldsymbol{\Sigma})d\boldsymbol{y} - \frac{2}{n}\sum_{i=1}^{n} f(\boldsymbol{y}_i|\boldsymbol{x}_i;\boldsymbol{B},\boldsymbol{\Sigma}) \tag{3}$$

where $\int f^2(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{B},\boldsymbol{\Sigma})d\boldsymbol{y} = 1/(2^q\pi^{q/2}|\boldsymbol{\Sigma}|^{1/2})$. Note that we assume a parametric family for the model while using a non-parametric ISE criterion to measure goodness of fit.

From the perspective of the loss function, ISE is a more robust measure of the goodness-of-fit compared with the likelihood-based loss function. It can match the model with the largest portion of the data because the integration in (3) accounts for the whole range of the squared loss function.

Using ISE criterion as the loss function, the objective function in (2) becomes

$$\begin{aligned}\tilde{L}_{n,\lambda}(\boldsymbol{B},\boldsymbol{\Sigma}) = \frac{|\boldsymbol{\Sigma}^{-1}|^{1/2}}{2^q\pi^{q/2}} - \frac{2}{n}\sum_{i=1}^{n} f(\boldsymbol{y}_i|\boldsymbol{x}_i;\boldsymbol{B},\boldsymbol{\Sigma}) \\ + \lambda_1\|\boldsymbol{\Sigma}^{-1}\|_1 + \lambda_2\|\boldsymbol{B}\|_1.\end{aligned} \tag{4}$$

However, the minimization of the objective in (4) is challenging. To circumvent this difficulty, we consider minimizing an upper bound of (4) which retains the robustness property.

For that purpose, we introduce a lemma which is essential for deriving the proposed objective (8) for estimating the sparse multiresponse regression model in (1).

LEMMA 1. *For a positive definite matrix $\boldsymbol{\Sigma}^{-1}$ with dimension q, the relation between its determinant value and l_1 norm can be described in the following inequality* $|\boldsymbol{\Sigma}^{-1}|^{1/2} \leq \left(\frac{\|\boldsymbol{\Sigma}^{-1}\|_1}{q}\right)^{q/2}$.

The proof of Lemma 1 is provided in the Appendix. Using Lemma 1, we can derive an upper bound for the objective function (4) as follows

$$c^*\|\boldsymbol{\Sigma}^{-1}\|_1^{q/2} - \frac{2}{n}\sum_{i=1}^{n} f(\boldsymbol{y}_i|\boldsymbol{x}_i;\boldsymbol{B},\boldsymbol{\Sigma}) + \lambda_1\|\boldsymbol{\Sigma}^{-1}\|_1 + \lambda_2\|\boldsymbol{B}\|_1, \tag{5}$$

where $c^* = 2^{-q}(\pi q)^{-q/2}$ is a constant. The above optimization problem amounts to minimizing

$$\breve{L}_{n,\lambda}(\boldsymbol{B},\boldsymbol{\Sigma}) = \breve{L}_n(\boldsymbol{B},\boldsymbol{\Sigma}) + \lambda_1^*\|\boldsymbol{\Sigma}^{-1}\|_1 + \lambda_2\|\boldsymbol{B}\|_1, \tag{6}$$

where $\breve{L}_n(\boldsymbol{B},\boldsymbol{\Sigma}) = -\frac{2}{n}\sum_{i=1}^{n} f(\boldsymbol{y}_i|\boldsymbol{x}_i;\boldsymbol{B},\boldsymbol{\Sigma})$ and λ_1^* is appropriately chosen. The value of λ_1^* here should be slightly larger than the value of λ_1 in (4). Moreover, the diagonal elements of $\boldsymbol{\Sigma}^{-1}$ are also penalized.

Note that $\breve{L}_{n,\lambda}(\boldsymbol{B},\boldsymbol{\Sigma})$ is an upper bound of $\tilde{L}_{n,\lambda}(\boldsymbol{B},\boldsymbol{\Sigma})$, however, the difference $\tilde{L}_{n,\lambda}(\boldsymbol{B},\boldsymbol{\Sigma}) - \breve{L}_{n,\lambda}(\boldsymbol{B},\boldsymbol{\Sigma})$ is well controlled by the penalty term $\lambda_1^*\|\boldsymbol{\Sigma}^{-1}\|_1$ in (6). By properly adjusting the value of λ_1^*, we can make the difference reasonably small. Therefore, by minimizing $\breve{L}_{n,\lambda}(\boldsymbol{B},\boldsymbol{\Sigma})$, we expect to approach the solution of $\tilde{L}_{n,\lambda}(\boldsymbol{B},\boldsymbol{\Sigma})$ and thus still retain the robustness property in the estimators.

Taking the logarithm on $\breve{L}_n(\boldsymbol{B},\boldsymbol{\Sigma})$, we obtain the loss

$$\begin{aligned}L_n(\boldsymbol{B},\boldsymbol{\Sigma}) = -\log\left[\frac{1}{n}\sum_{i=1}^{n}\exp(-\frac{1}{2}(\boldsymbol{y}_i-\boldsymbol{B}'\boldsymbol{x}_i)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i-\boldsymbol{B}'\boldsymbol{x}_i))\right] \\ -\frac{1}{2}\log|\boldsymbol{\Sigma}^{-1}|.\end{aligned} \tag{7}$$

We note that the logarithm is employed to strike a better balance between goodness of fit and the sparsity inducing penalty (similarly one considers the penalized negative log-likelihood rather than dealing with the likelihood directly). This yields the estimator proposed and studied in this paper, the Regularized Integrated Square Error (REG-ISE) estimator, which minimizes the following objective function:

$$\begin{aligned}&L_{n,\lambda}(\boldsymbol{B},\boldsymbol{\Sigma}) = \\ &-\log\left[\frac{1}{n}\sum_{i=1}^{n}\exp(-\frac{1}{2}(\boldsymbol{y}_i-\boldsymbol{B}'\boldsymbol{x}_i)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i-\boldsymbol{B}'\boldsymbol{x}_i))\right] \\ &-\frac{1}{2}\log|\boldsymbol{\Sigma}^{-1}| + \lambda_1\|\boldsymbol{\Sigma}^{-1}\|_1 + \lambda_2\|\boldsymbol{B}\|_1.\end{aligned} \tag{8}$$

For notational convenience, here and in later sections, we use λ_1 for λ_1^*.

Some intuition can already be gained on the objective robustness by considering the ratio between data and model pdf: $f(\boldsymbol{y}|\boldsymbol{x})/f(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{B},\boldsymbol{\Sigma})$. An outlier in the data may drives this ratio to infinity, in which case the log-likelihood is also infinity. In contrast, the difference $f(\boldsymbol{y}|\boldsymbol{x}) - f(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{B},\boldsymbol{\Sigma})$ is always bounded. This property makes the L_2-distance a favourable choice when dealing with outliers. We note that a similar reasoning holds in the context of density estimation, as pointed out in the recent work of [30] on density-difference estimation.

3.2 Optimization

The REG-ISE objective function (8) is non-convex and non-smooth. In order to solve it, one could consider approximating the "log-sum-exp" term in the objective, combined with alternate optimization for $\boldsymbol{B}$ and $\boldsymbol{\Sigma}^{-1}$ respectively (as done in MRCE). However, the convergence of alternate optimization can be very slow, as observed in the case of MRCE [27].

Instead, we propose to adopt a sequential quadratic programming algorithm recently developed by Curtis & Overton [9] for the non-smooth and non-convex optimization. The basic idea of their algorithm SLQP-GS is to combine sequential quadratic approximation with a process of gradient sampling so that the computation of the search direction is effective in nonsmooth regions. The only requirement for SLQP-GS to be applicable is that the objective and constraints (if any) be continuously differentiable on open dense subsets, which is satisfied in our case. We also benefit from the convergence guarantees of SLQP-GS, namely that the algorithm is guaranteed to converge to a solution regardless of the initialization with probability one.

We employed the Matlab implementation of SLQP-GS provided by the authors, which is available from `http://coral.ie.lehigh.edu/~frankecurtis/software`. Due to space constraints, we refer the reader to Curtis & Overton [9] for details on the algorithm, its matlab implementation and convergence results. We note that the gradient sampling step in SLQP-GS can be efficiently parallelized for fast computation in high dimensional applications. Alternatively one can perform adaptive sampling of gradients over the course of the optimization process as described in [10].

3.3 Consistency Results

The L_2 distance estimators are known to strike the right balance between statistical efficiency and robustness [2, 28]. In this section, we add to this body of evidence by showing that the REG-ISE estimator is root-n consistent for the settings of the fixed dimensionality p. Denote by $\bar{\boldsymbol{B}}$ the true regression coefficient matrix, and by $\bar{\boldsymbol{\Sigma}}$ the true covariance matrix. We assume the following conditions:
(C1) $\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} \to \boldsymbol{A}$, where $\boldsymbol{A}$ is positive definite.
(C2) There exist $\sqrt{n}$-consistent estimators of $\bar{\boldsymbol{B}}$ and $\bar{\boldsymbol{\Sigma}}^{-1}$.
Condition (C2) can be replaced by some technical regularity conditions as in [13] so as to guarantee the consistency of ordinary maximum likelihood estimators.

THEOREM 1. *Consider sequences $\lambda_{1,n}$ and $\lambda_{2,n}$ of regularization parameters, such that*

$$\lambda_{1,n} n^{-1/2} \to 0 \text{ and } \lambda_{2,n} n^{-1/2} \to 0.$$

Then under the conditions (C1) and (C2), there exists a local minimizer $(\hat{\boldsymbol{B}}, \hat{\boldsymbol{\Sigma}}^{-1})$ of REG-ISE such that

$$\|(vec(\hat{\boldsymbol{B}})', vec(\hat{\boldsymbol{\Sigma}}^{-1})')' - (vec(\bar{\boldsymbol{B}})', vec(\bar{\boldsymbol{\Sigma}}^{-1})')\| = O_p(1/\sqrt{n}).$$

The proof is provided in the Appendix.

The above theoretical results hold for the case where the dimensionality p is fixed, while the sample size n is allowed to grow. As a future work we plan to extend our results to the case where p is allowed to grow with the sample size n. We conjecture that condition (C1) might be replaced by conditions on sample and population covariance while (C2) is still achieved by certain penalized maximum likelihood estimators [23]. We note however that such an extension is highly non-trivial. In fact, to our knowledge, no theory is yet available for *joint* estimation in the high dimensional case even for the standard maximum likelihood estimator, the non-convexity being the source of the difficulty. Nevertheless, in view of our superior empirical results, we hope that theoretical results can be obtained by making use of the techniques in [26] which solely deal with inverse covariance estimation, and those in [23] which solely concern regression.

3.4 Insights into Robustness

We provide some insights for the robustness of REG-ISE by considering a first order approximation of the "log-sum-exp" term in (8).

Define the parameter set $\boldsymbol{\beta} = (\boldsymbol{B}, \boldsymbol{\Sigma}^{-1})$ and denote $g_i(\boldsymbol{\beta}) = -\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{B}'\boldsymbol{x}_i)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i - \boldsymbol{B}'\boldsymbol{x}_i)$. We consider a first-order approximation for $\log\left[\frac{1}{n}\sum_{i=1}^n \exp(g_i(\boldsymbol{\beta}))\right]$ with respect to $\boldsymbol{\beta}$ as follows:

$$\log\left[\frac{1}{n}\sum_{i=1}^n \exp(g_i(\boldsymbol{\beta}))\right] \approx$$
$$C_0 + \frac{1}{n}\sum_{i=1}^n \frac{\exp(g_i(\boldsymbol{\beta}_0))}{\frac{1}{n}\sum_{i=1}^n \exp(g_i(\boldsymbol{\beta}_0))} \nabla g_i(\boldsymbol{\beta}_0)^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0),$$

where $\boldsymbol{\beta}_0$ is an initial estimate and C_0 is some constant independent of $\boldsymbol{\beta}$.

Using the fact that $g_i(\boldsymbol{\beta}) \approx g_i(\boldsymbol{\beta}_0) + \nabla g_i(\boldsymbol{\beta}_0)^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$, we have the following

$$\log\left[\frac{1}{n}\sum_{i=1}^n \exp(g_i(\boldsymbol{\beta}))\right] \propto \frac{1}{n}\sum_{i=1}^n \frac{\exp(g_i(\boldsymbol{\beta}_0))}{\frac{1}{n}\sum_{i=1}^n \exp(g_i(\boldsymbol{\beta}_0))} g_i(\boldsymbol{\beta}),$$

up to some constant independent of $\boldsymbol{\beta}$. Therefore, the objection function (8) can be approximated by

$$-\log|\boldsymbol{\Sigma}^{-1}| + \frac{1}{n}\sum_{i=1}^n w_i(\boldsymbol{y}_i - \boldsymbol{B}'\boldsymbol{x}_i)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i - \boldsymbol{B}'\boldsymbol{x}_i)$$
$$+ \lambda_1\|\boldsymbol{\Sigma}^{-1}\|_1 + \lambda_2\|\boldsymbol{B}\|_1, \quad (9)$$

up to some constant and where

$$w_i \equiv w_i(\boldsymbol{\beta}_0) = \frac{\exp(g_i(\boldsymbol{\beta}_0))}{\frac{1}{n}\sum_{i=1}^n \exp(g_i(\boldsymbol{\beta}_0))}. \quad (10)$$

By defining $\boldsymbol{S}^* = \boldsymbol{S}^*(\boldsymbol{\beta}_0) = \frac{1}{n}\sum_{i=1}^n w_i(\boldsymbol{y}_i - \boldsymbol{B}'\boldsymbol{x}_i)(\boldsymbol{y}_i - \boldsymbol{B}'\boldsymbol{x}_i)'$, we can rewrite (9) as

$$-\log|\boldsymbol{\Sigma}^{-1}| + \text{trace}[\boldsymbol{\Sigma}^{-1}\boldsymbol{S}^*(\boldsymbol{\beta}_0)] + \lambda_1\|\boldsymbol{\Sigma}^{-1}\|_1 + \lambda_2\|\boldsymbol{B}\|_1. \quad (11)$$

Note that $\boldsymbol{S}^*$ can be viewed as a weighted sample covariance matrix where weights are with respect to n observations.

One could then envision an approximate iterative procedure where given initial estimates, data are first re-weighted by w_i in (10) and then alternately passed to lasso and l_1 penalized inverse covariance solvers (e.g. QUIC, glasso) to provide new estimates, and the procedure would be repeated until convergence (see details in the appendix). This intuitively elaborates the robustness property of REG-ISE. Indeed the weights w_i are proportional to the likelihood functions of individual data points, i.e., $w_i = \frac{L(\boldsymbol{y}_i|\boldsymbol{x}_i;\beta_0)}{\sum_{i=1}^n L(\boldsymbol{y}_i|\boldsymbol{x}_i;\beta_0)}$. Thus data with high likelihood values are given more weights in the estimation. Conversely, data with low likelihood values, which are more likely to be outliers, contribute less to the estimation. The connection between the likelihood functions and weights nicely explains the resilience of the proposed estimator to outliers.

3.5 Tuning Parameter Selection

Approaches for choosing tuning parameters include cross-validation (CV) [4], the hold-out validation set method [21],

and information criteria such as Bayesian information criterion (BIC) [34]. Here we proposed a modified scheme for the cross-validation method. The common K-fold CV consists in randomly partitioning the data into K folds, and then leaving out one fold of data as validation set while all the other folds are used as training set in each CV iteration. Note that CV assumes that the data are i.i.d. distributed, and therefore the validation set and training set are considered statistically equivalent. However, such an assumption is no longer valid in the presence of outliers since the proportions of the outliers in the validation data and training data can be different. Consequently, the validation set cannot be used to evaluate the model obtained by the training set.

To tackle this issue, we develop a modified cross-validation scheme motivated by the idea of sliced designs [25]. Specifically, we perform K-fold cross-validation for $n = mK$ observations as follows. Based on initial estimates of the model parameters, we first rank the observed data according to the values of their likelihood functions. Then the first K data points are randomly assigned to K folds, one point per fold. Subsequently the next K data points are randomly assigned to K folds. This procedure is repeated m times. In this way, the data in each fold are more likely to have similar distributions. This modified scheme can also be applied to tuning via hold-out validation set method.

4. SIMULATION STUDY

We compare the proposed REG-ISE with MRCE [27], l_1-CGGM [29], CAPME [7], and LAD-Lasso [32]. Note that LAD-Lasso can only estimate regression coefficients.

In our experiments, the rows of $n \times p$ predictor matrices $\boldsymbol{X}$ are sampled independently from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma_x})$ where $(\boldsymbol{\Sigma_x})_{i,j} = 0.5^{|i-j|}$. We consider the following two cases.

Case 1: The covariance matrix is set to $\boldsymbol{\Sigma}_{i,j} = 0.7^{|i-j|}$ which corresponds to an AR(1) model with banded $\boldsymbol{\Sigma}^{-1}$. We randomly select 10 percent of the predictors to be irrelevant to all the responses. Then for each response, we randomly select half of the remaining predictors to be relevant for that response. The corresponding non-zero entries in the regression matrix $\boldsymbol{B}$ are sampled independently from $\mathcal{N}(0, 1)$. We consider 60 predictors, 20 responses and 100 observations.

Case 2: $\boldsymbol{\Sigma}^{-1}$ is the graph Laplacian of a tree with outdegree of 4 and edge weights uniformly sampled from $[0.3, 1.0]$. For each response we randomly select 10 percent of the predictors to be relevant, and sample the corresponding nonzero enties in $\boldsymbol{B}$ independently from $\mathcal{N}(0, 1)$. We consider 1000 predictors, 100 responses and 400 observations.

To address the robustness issue, we consider various percentages of outliers contaminating the responses. The uncontaminated data are generated from $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{B}'\boldsymbol{x}, \boldsymbol{\Sigma})$, where $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$ are specified above. Two scenarios presenting outliers are considered: (i) outliers with respect to the mean: $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{B}'\boldsymbol{x} + \boldsymbol{C}, \boldsymbol{\Sigma})$ where $\boldsymbol{C}$ is a constant vector of 5 and (ii) outliers regarding the covariance structure: $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{B}'\boldsymbol{x}, \boldsymbol{I})$ where $\boldsymbol{I}$ is an identity matrix.

To measure variable selection accuracy, we use the F_1 score defined by $F_1 = 2PR/(P + R)$, where P is precision (fraction of correctly selected variables among selected variables) and R is recall (fraction of correctly selected variables among true relevant variables). To measure the estimation accuracy of $\boldsymbol{B}$, we report the model error defined as $ME(\hat{\boldsymbol{B}}, \boldsymbol{B}) = \text{tr}\left[(\hat{\boldsymbol{B}} - \boldsymbol{B})^T \boldsymbol{\Sigma_x} (\hat{\boldsymbol{B}} - \boldsymbol{B})\right]$, where $\hat{\boldsymbol{B}}$ is the estimated regression coefficient matrix. The estimation accuracy for $\boldsymbol{\Sigma}^{-1}$ is measured by its l_2 loss, defined as $\|\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\|_F$, under Frobenius norm where $\hat{\boldsymbol{\Sigma}}^{-1}$ is the estimated inverse covariance matrix.

For each of the above settings, we generate 50 simulated dataset. For all five comparison methods, for each dataset we use the modified 5-fold cross-validation described in Section 3.5 to tune parameters λ_1 and λ_2.

The choice of initial parameter estimates is important for MRCE and REG-ISE as their respective objective functions are non-convex. The initial estimates for $\boldsymbol{B}$ are obtained using ridge regression. In addition for REG-ISE, $\boldsymbol{\Sigma}$ is initialized as the inverse of the sample covariance matrix of ridge regression residuals perturbed by a positive diagonal matrix.

Figures 1 and 2 present the results for Case 1. Results for Case 2 are summarized in Table 1 Similar behavior in terms of robustness is observed for both cases. Our proposed REG-ISE estimator clearly outperforms other methods due to its robustness against outliers with respect to both mean and covariance deviations. Performance of MRCE, l_1-CGGM and CAPME seriously degrade once outliers are introduced. Surprisingly, LAD-Lasso does not show much resilience to outliers. Moreover, with "clean" data, its estimation and variable selection accuracy is inferior to other methods as it ignores the dependencies among responses. Interestingly, even when there are no outliers in the data, REG-ISE is competitive, since it is more likely to distinguish true signals from various noise amplitudes.

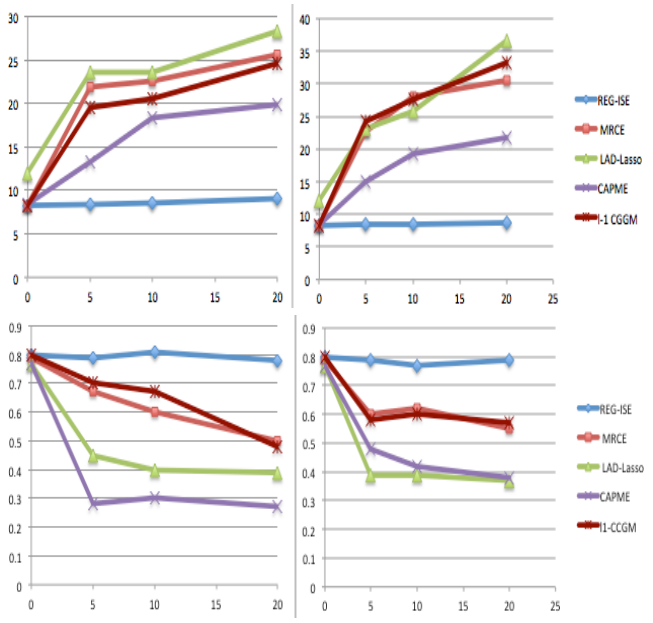

Figure 1: Average model error $ME(\hat{\boldsymbol{B}}, \boldsymbol{B})$ (top) and F_1 scores (bottom) for $\boldsymbol{B}$ estimated by REG-ISE, MRCE, l_1-CGGM, LAD-Lasso, and CAPME on simulated data of Case 1. Outliers in terms of the mean (left), and covariance (right). The x-axis corresponds to the percentage of outliers.

Table 1: Simulation results for Case 2. Top: Model error for B/l_2 loss for Σ^{-1}. Bottom: F_1 for B/F_1 for Σ^{-1}.

Measure	Outlier Type	Outlier %	REG-ISE	MRCE	l-1 CGGM	CAPME	LAD-Lasso
ME(B)/$l_2(\Sigma^{-1})$	None	0	33.12/48.3	33.92/61.88	32/47.6	33.4/49.6	47.96/NA
ME(B)/$l_2(\Sigma^{-1})$	Mean	5	33.44/49.28	87.56/109.4	78/75.8	53.12/59.3	94.04/NA
ME(B)/$l_2(\Sigma^{-1})$	Mean	10	34/52.84	90.28 / 178.2	82.4/80	73.2/72.79	94.24/NA
ME(B)/$l_2(\Sigma^{-1})$	Mean	20	35.96/53.4	102.6/196.3	98.4/101	79.36/80.23	113/NA
ME(B)/$l_2(\Sigma^{-1})$	Cov	5	34.04/51.04	90.96/81.04	97.36/62.8	60.2/96.34	92.68/NA
ME(B)/$l_2(\Sigma^{-1})$	Cov	10	34.29/52.6	96.39/102.1	110.16/104.2	77.28/156.2	103.6/NA
ME(B)/$l_2(\Sigma^{-1})$	Cov	20	36.5/56.92	122.6/186.4	133/167.1	87.4/153.4	146.68/NA
F_1(B)/$F_1(\Sigma^{-1})$	None	0	0.8/0.77	0.71/0.32	0.79/0.78	0.72/0.76	0.6/NA
F_1(B)/$F_1(\Sigma^{-1})$	Mean	5	0.78/0.79	0.65/0.24	0.65/0.7	0.28/0.59	0.45/NA
F_1(B)/$F_1(\Sigma^{-1})$	Mean	10	0.78/0.76	0.54/0.23	0.67/0.58	0.3/0.57	0.4/NA
F_1(B)/$F_1(\Sigma^{-1})$	Mean	20	0.76/0.75	0.49/0.21	0.48/0.49	0.27/0.57	0.39/NA
F_1(B)/$F_1(\Sigma^{-1})$	Cov	5	0.77/0.75	0.58/0.3	0.59/ 0.58	0.48/0.68	0.39/NA
F_1(B)/$F_1(\Sigma^{-1})$	Cov	10	0.77/0.75	0.64/0.29	0.57/0.57	0.42/0.67	0.39/NA
F_1(B)/$F_1(\Sigma^{-1})$	Cov	20	0.78/0.73	0.38/0.23	0.49/0.47	0.38/0.48	0.37/NA

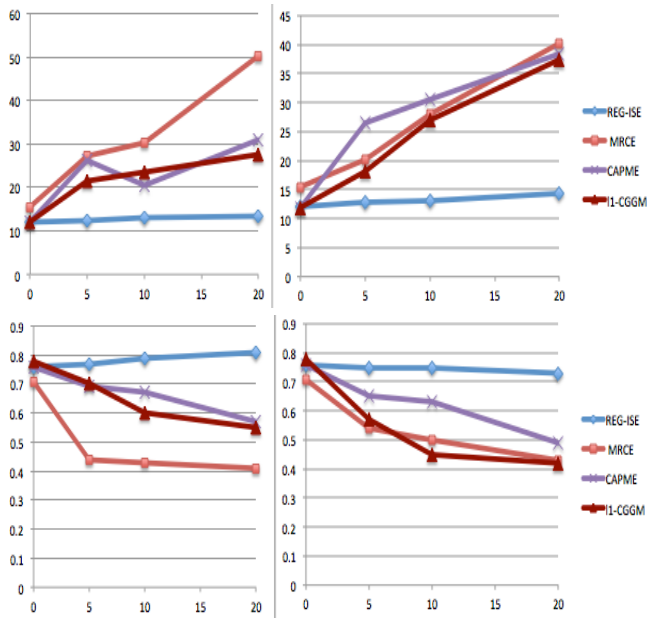

Figure 2: Average estimation error $\|\hat{\Sigma}^{-1} - \Sigma^{-1}\|_F$ (top) and F_1 scores (bottom) for Σ^{-1} estimated by REG-ISE, MRCE, l_1-CGGM, LAD-Lasso, and CAPME on simulated data of Case 1. Outliers in terms of the mean (left), and covariance (right). The x-axis corresponds to the percentage of outliers.

5. APPLICATIONS

In this section, we illustrate the usefulness of the proposed robust methods through two motivating applications and compare the results of our robust estimators with those of existing methods.

5.1 Asset Return Prediction

As a toy example for multivariate time series, we analyze a financial dataset which has been studied in [27] and [34]. This dataset contains weekly log-returns of 9 stocks for year 2004. Given multivariate time series data of log-returns $\boldsymbol{y}_t$ for weeks $t = 1, ..., T$, a first-order vector autoregressive model is considered as follows

$$\boldsymbol{y}_t = \boldsymbol{B}'\boldsymbol{y}_{t-1} + \boldsymbol{\epsilon}_t, \boldsymbol{\epsilon}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), t = 2, \ldots, T$$

where the response matrix $\boldsymbol{y}_t$ is formed by observations at week t and the predictor matrix $\boldsymbol{y}_{t-1}$ contains observations at the previous week $t-1$. Following the analysis in Rothman et al. [27], we used log-returns of the first 26 weeks as training set, and log-returns of the remaining 26 weeks as testing set. The tuning parameters were selected using the modified 10-fold cross-validation described in Section 2.4. Table 2 reports the mean squared prediction error (MSPE) of the five comparison methods. Even though all methods are competitive on this dataset, REG-ISE estimator achieves the smallest prediction error. Figure 3 presents the graphs induced by the estimates of $\boldsymbol{\Sigma}^{-1}$ using MRCE and REG-ISE, respectively. Comparing the two graphs, both MRCE and REG-ISE indicate that companies from the same industry are partially correlated, e.g. GE and IBM (technology), Ford and GM (auto industry). AIG (insurance company) seems to be partially correlated with most of the other companies. However, there are some discrepancies between the two graphs, e.g., GM is found to be partially correlated to IBM by REG-ISE but to be uncorrelated by MRCE. Overall, the results from REG-ISE have reasonable financial interpretation.

5.2 eQTL Data Analysis

We analyze yeast eQTL dataset [6] which contains genotype data for 2,956 SNPs (predictors) and microarray data for 6,216 genes (responses) regarding 112 segregants (instances). We extracted 1,260 unique SNPs, and focused on 125 genes belonging to cell-cycle pathway provided by the KEGG database [18]. For all methods, the tuning parameters were chosen via 5-fold modified cross-validation de-

Table 2: Prediction accuracy measured by MSPE for various methods for the asset return dataset.

Method	MSPE
REG-ISE	$\mathbf{0.69 \pm 0.11}$
MRCE	0.71 ± 0.12
l_1-CGGM	0.72 ± 0.10
CAPME	0.72 ± 0.11
LAD-Lasso	0.73 ± 0.12

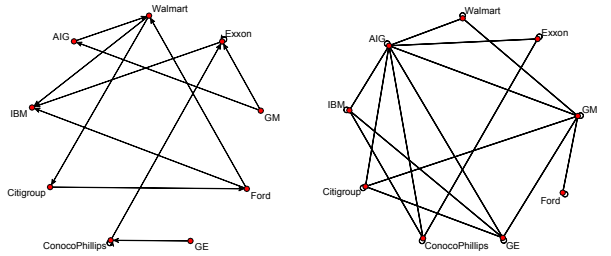

Figure 3: Graphs from the estimates of the inverse covariance matrix $\mathbf{\Sigma}^{-1}$ obtained by MRCE (left) and REG-ISE(right).

scribed in Section 3.5. We first evaluated the predictive accuracy of our method and the comparison methods by randomly partitioning the data into training and test sets, using 90 observations for training and the remainder for testing. We computed the MSPE for the testing set. The average MSPEs based on 20 random partitions are presented in Table 3. We can see that overall the predictive performance of REG-ISE is superior to the other methods.

Figure 4 shows the cell-cycle pathways estimated by the proposed REG-ISE, MRCE method, CAPME method along with the benchmark KEGG pathway. MRCE tends to estimate many spurious links. Similar observation holds for l_1-CGGM, so its estimated graph is not represented here. CAPME recovers some of the links but not as accurately as REG-ISE. This can be partly explained by the fact that CAPME does not take into account the covariance structure in its regression stage and does not have any feedback loop. This can result in poor estimation of the regression matrix $\boldsymbol{B}$, which in turn may negatively impact the estimation of precision matrix $\mathbf{\Sigma}^{-1}$. In addition, lack of robustness can also result in inaccurate network reconstruction. Certain discrepancies between true and estimated graphs may also be caused by inherent limitations in this dataset. For instance, some edges in cell-cycle pathway may not be observable from gene expression data. Additionally in this dataset, the perturbation of cellular systems may not be significant enough to enable accurate inference of some of the links.

Using the KEGG pathway as the "ground truth", we also computed the F_1 scores for the estimates of $\mathbf{\Sigma}^{-1}$ shown in Table 4. As a sanity check, we analyzed the microarray data without SNPs as predictors using glasso. The resulting graph was extremely dense with F_1 score to be 0.033. This indicates the disadvantage of procedures like glasso which is unable to adjust for predictors (hereby the genetic variants) in inverse covariance matrix estimation.

Table 3: MSPEs under different methods based on 20 random partitions of the eQTL into training and test sets.

REG-ISE	MRCE	l_1-CGGM	CAPME
$\mathbf{2.36 \pm 0.07}$	6.25 ± 0.22	4.46 ± 0.17	4.38 ± 0.09

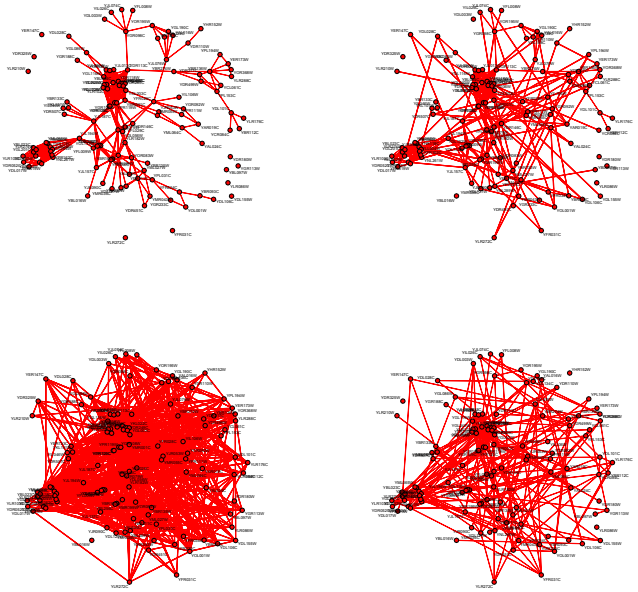

Figure 4: Yeast cell cycle network provided in the KEGG database (top left), estimated by REG-ISE (top right), MRCE (bottom left), and CAPME (bottom right).

Table 4: F_1 scores of the estimated cell-cycle network (higher values indicate higher accuracy).

REG-ISE	MRCE	l_1-CGGM	CAPME	Glasso
0.635 $\pm$0.009	0.042 $\pm$0.008	0.089 $\pm$0.011	0.348 $\pm$0.034	0.033 $\pm$0.014

From reconstructed network and F_1 scores, we conclude that REG-ISE most faithfully estimates the cell-cycle network compared to the other methods, which clearly demonstrates the value of embracing the robustness.

6. CONCLUDING REMARKS

In this work, we have developed a robust framework to jointly estimate multiresponse regression and inverse covariance matrix from high dimensional data. Our formulation is readily applicable to deal with single regression and sparse inverse covariance estimation by itself, as well as to the parameterization used in [29], which will be an interesting future work. The proposed methodology is valuable for many applications beyond the integration of genomic and transcriptomic data and finantial data analysis. Additional interesting future work includes extending the proposed method to directed graph modeling via vector autoregressive models, and extending our theoretical results to the high-dimensional setting.

APPENDIX

A. PROOF OF LEMMA 1

Suppose that the eigenvalues of $\mathbf{\Sigma}^{-1}$ are $d_i, i = 1, \ldots, q$. Using the fact that $\sqrt[q]{\prod_{i=1}^{q} d_i} \leq \frac{1}{q}\sum_{i=1}^{q} d_i$, one can have

$$|\mathbf{\Sigma}^{-1}|^{1/2} \leq \left(\frac{\sum_{i=1}^{q} d_i}{q}\right)^{q/2}.$$

For each eigenvalue d_i, we apply the Gershgorin's circle theorem to obtain its upper bound. That is

$$|d_i - c_{ii}| \leq \sum_{j\neq i} |c_{ij}| \Rightarrow d_i \leq \sum_{j=1}^{q} |c_{ij}|$$

where $c_{ij}, i = 1, \ldots, q; j = 1, \ldots, q$ are the elements in matrix $\mathbf{\Sigma}^{-1}$. Therefore, we have $\sum_{i=1}^{q} d_i \leq \sum_{i=1}^{q}\sum_{j=1}^{q} |c_{ij}| = \|\mathbf{\Sigma}^{-1}\|_1$, which leads to

$$|\mathbf{\Sigma}^{-1}|^{1/2} \leq \left(\frac{\|\mathbf{\Sigma}^{-1}\|_1}{q}\right)^{q/2}.$$

This ends the proof of Lemma 1.

B. PROOF SKETCH OF THEOREM 1

Recall our objective function:

$$L_{n,\lambda}(\boldsymbol{B}, \mathbf{\Sigma}^{-1}) = - \log\left[\frac{1}{n}\sum_{i=1}^{n} \exp(-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{B}'\boldsymbol{x}_i)'\mathbf{\Sigma}^{-1}(\boldsymbol{y}_i - \boldsymbol{B}'\boldsymbol{x}_i))\right] - \frac{1}{2}\log|\mathbf{\Sigma}^{-1}| + \lambda_1\|\mathbf{\Sigma}^{-1}\|_1 + \lambda_2\|\boldsymbol{B}\|_1 \quad (12)$$

Let $\bar{\boldsymbol{B}}$ and $\bar{\mathbf{\Sigma}}^{-1}$ be the true regression and inverse covariance matrices. We follow the same reasoning as in the proof of Theorem 1 in [13]. The key idea is that it's enough to show that for any $\delta > 0$ there exists a large constant C, such that

$$P\{\sup_{\|\boldsymbol{U}\|=C} L_{n,\lambda}(\bar{\boldsymbol{B}} + \frac{\boldsymbol{U}_1}{\sqrt{n}}, \bar{\mathbf{\Sigma}}^{-1} + \frac{\boldsymbol{U}_2}{\sqrt{n}}) > L_{n,\lambda}(\bar{\boldsymbol{B}}, \bar{\mathbf{\Sigma}}^{-1})\} \geq 1 - \delta$$

with $\boldsymbol{U} = (vec(\boldsymbol{U}_1)', vec(\boldsymbol{U}_2)')'$.
Define $Q(\boldsymbol{B}, \mathbf{\Sigma}^{-1})$ as

$$Q(\boldsymbol{B}, \mathbf{\Sigma}^{-1}) = -\log\left[\frac{1}{n}\sum_{i=1}^{n} \exp(-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{B}'\boldsymbol{x}_i)'\mathbf{\Sigma}^{-1}(\boldsymbol{y}_i - \boldsymbol{B}'\boldsymbol{x}_i))\right].$$

By using a similar reasoning as in Section 3.4, one can show that around the true parameters $(\bar{\boldsymbol{B}}, \bar{\mathbf{\Sigma}}^{-1})$, the difference $Q(\bar{\boldsymbol{B}} + \frac{\boldsymbol{U}_1}{\sqrt{n}}, \bar{\mathbf{\Sigma}}^{-1} + \frac{\boldsymbol{U}_2}{\sqrt{n}}) - Q(\bar{\boldsymbol{B}}, \bar{\mathbf{\Sigma}}^{-1})$ can be lower-bounded by

$$\frac{1}{n}\sum_{i=1}^{n} w_i(\boldsymbol{y}_i - (\bar{\boldsymbol{B}} + \frac{\boldsymbol{U}_1}{\sqrt{n}})'\boldsymbol{x}_i)'(\bar{\mathbf{\Sigma}}^{-1} + \frac{\boldsymbol{U}_2}{\sqrt{n}})(\boldsymbol{y}_i - (\bar{\boldsymbol{B}} + \frac{\boldsymbol{U}_1}{\sqrt{n}})'\boldsymbol{x}_i) - \frac{1}{n}\sum_{i=1}^{n} w_i(\boldsymbol{y}_i - \bar{\boldsymbol{B}}'\boldsymbol{x}_i)'\mathbf{\Sigma}^{-1}(\boldsymbol{y}_i - \bar{\boldsymbol{B}}'\boldsymbol{x}_i) + o(1)$$

where $w_i \equiv w_i(\boldsymbol{\beta}_0) = \frac{\exp(g_i(\boldsymbol{\beta}_0))}{\frac{1}{n}\sum_{i=1}^{n}\exp(g_i(\boldsymbol{\beta}_0))}$. Even though Q is globally non-convex, such an approximation is valid as Q is locally bi-convex in a neighborhood of the true parameters $(\bar{\boldsymbol{B}}, \bar{\mathbf{\Sigma}}^{-1})$ with asymptotic probability one. Briefly, one can show that the events $\frac{\boldsymbol{U}_1^T}{\sqrt{n}}\nabla_B^2 Q(\bar{\boldsymbol{B}} + t\frac{\boldsymbol{U}_1}{\sqrt{n}})\frac{\boldsymbol{U}_1}{\sqrt{n}} < 0$ and $\frac{\boldsymbol{U}_2^T}{\sqrt{n}}\nabla_{\Sigma^{-1}}^2 Q(\bar{\mathbf{\Sigma}}^{-1} + t\frac{\boldsymbol{U}_2}{\sqrt{n}})\frac{\boldsymbol{U}_2}{\sqrt{n}} < 0$ for $t \in (0, 1)$ are small uniformly on $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ as long as $\frac{\|\boldsymbol{U}\|}{\sqrt{n}}$ is small enough. Briefly, this can be done by 'brute force' calculation of the Hessians around the true solution, noticing that the expected value of $\frac{\boldsymbol{U}_1^T}{\sqrt{n}}\nabla_B^2 Q(\bar{\boldsymbol{B}} + t\frac{\boldsymbol{U}_1}{\sqrt{n}})\frac{\boldsymbol{U}_1}{\sqrt{n}}$ and $\frac{\boldsymbol{U}_2^T}{\sqrt{n}}\nabla_{\Sigma^{-1}}^2 Q(\bar{\mathbf{\Sigma}}^{-1} + t\frac{\boldsymbol{U}_2}{\sqrt{n}})\frac{\boldsymbol{U}_2}{\sqrt{n}}$ are both non-negative for $\frac{\|\boldsymbol{U}\|}{\sqrt{n}}$ small enough and then upper-bounding the large deviations from the expected values using Azuma-Hoeffding's inequality.

Let us define $V_n(\boldsymbol{U})$ as

$$V_n(\boldsymbol{U}) = L_{n,\lambda}(\bar{\boldsymbol{B}} + \frac{\boldsymbol{U}_1}{\sqrt{n}}, \bar{\mathbf{\Sigma}}^{-1} + \frac{\boldsymbol{U}_2}{\sqrt{n}}) - L_{n,\lambda}(\bar{\boldsymbol{B}}, \bar{\mathbf{\Sigma}}^{-1}).$$

For notation convenience, denote $\tilde{\boldsymbol{Y}} = \mathrm{diag}(\sqrt{w_1}, \ldots, \sqrt{w_n})\boldsymbol{Y}$, and $\tilde{\boldsymbol{X}} = \mathrm{diag}(\sqrt{w_1}, \ldots, \sqrt{w_n})\boldsymbol{X}$. Noting that $|\bar{\boldsymbol{B}}_{jk} + \frac{u_{1j,k}}{\sqrt{n}}| - |\bar{\boldsymbol{B}}_{jk}| = |\frac{u_{1j,k}}{\sqrt{n}}|$ for $\bar{\boldsymbol{B}}_{jk} = 0$ and $|\bar{\mathbf{\Sigma}}_{st}^{-1} + \frac{u_{2s,t}}{\sqrt{n}}| - |\bar{\mathbf{\Sigma}}_{st}^{-1}| = |\frac{u_{2s,t}}{\sqrt{n}}|$ for $\bar{\mathbf{\Sigma}}_{st}^{-1} = 0$, we then have

$$\begin{aligned} V_n(\boldsymbol{U}) \geq & -\log|(\bar{\mathbf{\Sigma}}^{-1} + \frac{\boldsymbol{U}_2}{\sqrt{n}})\bar{\mathbf{\Sigma}}| \\ & + \frac{1}{n}\mathrm{trace}\{(\bar{\mathbf{\Sigma}}^{-1} + \frac{\boldsymbol{U}_2}{\sqrt{n}})(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}(\bar{\boldsymbol{B}} + \frac{\boldsymbol{U}_1}{\sqrt{n}}))'(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}(\bar{\boldsymbol{B}} + \frac{\boldsymbol{U}_1}{\sqrt{n}}))\} \\ & - \frac{1}{n}\mathrm{trace}\{\bar{\mathbf{\Sigma}}^{-1}(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\bar{\boldsymbol{B}})'(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\bar{\boldsymbol{B}})\} \\ & + \lambda_1 \sum_{\boldsymbol{B}_{kj}\neq 0} (|\bar{\boldsymbol{B}}_{jk} + \frac{u_{1j,k}}{\sqrt{n}}| - |\bar{\boldsymbol{B}}_{jk}|) \\ & + \lambda_2 \sum_{\bar{\mathbf{\Sigma}}_{st}^{-1}\neq 0} (|\bar{\mathbf{\Sigma}}_{st}^{-1} + \frac{u_{2s,t}}{\sqrt{n}}| - |\bar{\mathbf{\Sigma}}_{jk}^{-1}|). \end{aligned}$$

Following the same derivations as the proof of Lemma 3 in [20], we can show that for a sufficiently large C, $V_n(\boldsymbol{U}) > 0$ uniformly on $\{\boldsymbol{U} : \|\boldsymbol{U}\| = C\}$ with probability greater than $1 - \delta$. This completes the proof.

C. APPROXIMATE ITERATIVE PROCEDURE

For completeness we provide the details of the approximate iterative procedure discussed in Section 3.4.

Algorithm 1 Approximate Iterative Procedure

Step 1: Given an initial estimate $\mathbf{\Sigma}_0^{-1}$ and an initial estimate $\boldsymbol{B}_0$.
Step 2: Compute w_i based on (10) and obtain $\boldsymbol{S}^* = \frac{1}{n}\sum_{i=1}^{n} w_i(\boldsymbol{y}_i - \boldsymbol{B}_0'\boldsymbol{x}_i)(\boldsymbol{y}_i - \boldsymbol{B}_0'\boldsymbol{x}_i)'$.
Step 3: Estimate $\mathbf{\Sigma}^{-1}$ by minimizing (11) given $\boldsymbol{B}_0$:

$$\hat{\mathbf{\Sigma}}^{-1} = \arg\min -\log|\mathbf{\Sigma}^{-1}| + \mathrm{trace}[\mathbf{\Sigma}^{-1}\boldsymbol{S}^*] + \lambda_1\|\mathbf{\Sigma}^{-1}\|_1.$$

Step 4: Estimate $\boldsymbol{B}$ by minimizing (9) given $\mathbf{\Sigma}_0^{-1}$:

$$\hat{\boldsymbol{B}} = \arg\min \frac{1}{n}\sum_{i=1}^{n} w_i(\boldsymbol{y}_i - \boldsymbol{B}'\boldsymbol{x}_i)'\mathbf{\Sigma}_0^{-1}(\boldsymbol{y}_i - \boldsymbol{B}'\boldsymbol{x}_i) + \lambda_2\|\boldsymbol{B}\|_1.$$

Step 5: If $\|\hat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}_0^{-1}\|_F^2 \leq \delta_1$ and $\|\hat{\boldsymbol{B}} - \boldsymbol{B}_0\|_F^2 \leq \delta_2$, stop. Else set $\mathbf{\Sigma}_0^{-1} = \hat{\mathbf{\Sigma}}^{-1}$ and $\boldsymbol{B}_0 = \hat{\boldsymbol{B}}$ and go back to Step 2.

References

[1] O. Banerjee, L. Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation. *JMLR*, 9:485–516, 2008.

[2] A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85, 1998.

[3] R. Beran. Robust location estimates. *Annals of Statistics*, 5:431–444, 1977.

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

[5] L. Breiman and J. H. Friedman. Predicting multivariate responses in multiple linear regression. *JRSS Series B.*, 1997.

[6] R. Brem and L. Kruglyak. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Natl. Acad. Sci. USA*, 102(5):1572–1577, 2005.

[7] T. T. Cai, H. Li, W. Liu, and J. Xie. Covariate adjusted precision matrix estimation with an application in genetical genomics. *Biometrika*, 2011.

[8] E. Candes and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35:2313–2351, 2007.

[9] F. E. Curtis and M. L. Overton. A Sequential Quadratic Programming Algorithm for Nonconvex, Nonsmooth Constrained Optimization. *SIAM Journal on Optimization*, 22(2):474–500, 2011.

[10] F. E. Curtis and X. Que. An Adaptive Gradient Sampling Algorithm for Nonsmooth Optimization. *Optimization Methods and Software*, 2011.

[11] D. L. Donoho and R. C. Liu. The "automatic" robustness of minimum distance functional. *Annals of Statistics*, 16:552–586, 1994.

[12] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *JASA*, 96:1348–1360, 2001.

[13] M. Finegold and M. Drton. Robust graphical modeling of gene networks using classical and alternative t-distribution. *Annals of Applied Statistics*, 5(2A):1075–1080, 2011.

[14] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[15] C. Hsieh, M. Sustik, I. Dhillon, and P. Ravikumar. Sparse inverse covariance matrix estimation using quadratic approximation. In *NIPS*, 2011.

[16] J. Huang, N. Liu, M. Pourahmadi, and L. Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98, 2006.

[17] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa. Kegg for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acid Research*, 31(1):355–360, 2010.

[18] S. L. Lauritzen. *Graphical Models*. Oxford: Clarendon Press, 1996.

[19] W. Lee and Y. Liu. Simultaneous multiple response regression and inverse covariance matrix estimation via penalized gaussian maximum likelihood. *J. Multivar. Anal.*, 2012.

[20] E. Levina, A. J. Rothman, and J. Zhu. Sparse estimation of large covariance matrices via a nested lasso penalty. *Annals of Applied Statistics*, 2(1), 2008.

[21] N. Meinshausen and P. Buhlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

[22] S. Negahban, P. Ravikumar, M. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *CoRR*, abs/1010.2731, 2010.

[23] G. Obozinski, B. Taskar, and M. Jordan. Multi-task feature selection. Tech. report, University of California, Berkeley, 2006.

[24] P. Z. G. Qian and C. F. J. Wu. Sliced space-filling designs. *Biometrika*, 96(4):945–956, 2009.

[25] P. Ravikumar, M. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. *Electron. J. Statist.*, 5:935–980, 2011.

[26] A. Rothman, E. Levina, and J. Zhu. Sparse multivariate regression with covariance estimation. *JCGS*, 19(4):947–962, 2010.

[27] D. Scott. Parametric statistical modeling by minimum integrated square error. *Technometrics*, 43:274–285, 2001.

[28] K. Sohn and S. Kim. Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization. *AISTATS*, 2012.

[29] M. Sugiyama, T. Suzuki, T. Kanamori, M. C. Du Plessis, S. Liu, and I. Takeuchi. Density-difference estimation. *NIPS*, 25:692–700, 2012.

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

[31] H. Wang, G. Li, and G. Jiang. Robust regression shrinkage and consistent variable selection through the LAD-lasso. *J. Business and Economics Statistics*, 25:347–355, 2007.

[32] J. Wolfowitz. The minimum distance method. *Annals of Mathematical Statistics*, 28(1):75–88, 1957.

[33] M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.