Integrative Multi-View Reduced-Rank Regression: Bridging Group-Sparse and Low-Rank Models

Gen Li[1], Xiaokang Liu[2], and Kun Chen[*2,3]

[1]Department of Biostatistics, Columbia University
[2]Department of Statistics, University of Connecticut, Storrs, CT
[3]Center for Population Health, University of Connecticut Health Center, Farmington, CT

Abstract

Multi-view data have been routinely collected in various fields of science and engineering. A general problem is to study the predictive association between multivariate responses and multi-view predictor sets, all of which can be of high dimensionality. It is likely that only a few views are relevant to prediction, and the predictors within each relevant view contribute to the prediction collectively rather than sparsely. We cast this new problem under the familiar multivariate regression framework and propose an *integrative reduced-rank regression* (iRRR), where each view has its own low-rank coefficient matrix. As such, latent features are extracted from each view in a supervised fashion. For model estimation, we develop a convex *composite nuclear norm penalization* approach, which admits an efficient algorithm via alternating direction method of multipliers. Extensions to non-Gaussian and incomplete data are discussed. Theoretically, we derive non-asymptotic oracle bounds of iRRR under a restricted eigenvalue condition. Our results recover oracle bounds of several special cases of iRRR including Lasso, group Lasso and nuclear norm penalized regression. Therefore, iRRR *seamlessly bridges group-sparse and low-rank methods* and can achieve substantially faster convergence rate under realistic settings of multi-view learning. Simulation studies and an application in the Longitudinal Studies of Aging further showcase the efficacy of the proposed methods.

Keywords: composite penalization; group selection; integrative multivariate analysis; multi-view learning; nuclear norm penalization.

---

[*]Corresponding author. kun.chen@uconn.edu.

1 Introduction

Multi-view data, or measurements of several distinct yet interrelated sets of characteristics pertaining to the same set of subjects, have become increasingly common in various fields. In a human lung study, for example, segmental airway tree measurements from CT-scanned images, patient behavioral data from questionnaires, gene expressions data, together with multiple pulmonary function test results from spirometry, were all collected. Unveiling lung disease mechanisms then amounts to linking the microscopic lung airway structures, the genetic information, and the patient behaviors to the global measurements of lung functions (Chen et al., 2016). In an Internet network analysis, the popularity and influence of a web page are related to its layouts, images, texts, and hyperlinks as well as by the content of other web pages that link back to it. In Longitudinal Study of Aging (LSOA) (Stanziano et al., 2010), the interest is to predict current health conditions of patients using historical information of their living conditions, household structures, habits, activities, medical conditions, among others. The availability of such multi-view data has made tackling many fundamental problems possible through an *integrative statistical learning* paradigm, whose success owes to the utilization of information from various lenses and angles simultaneously.

The aforementioned problems can all be cast under a multivariate regression framework, in which both the responses and the predictors can be high dimensional, and in addition, the predictors admit some natural grouping structure. In this paper we investigate this simple yet general framework for achieving integrative learning. To formulate, suppose we observe $\mathbf{X}_k \in \mathbb{R}^{n \times p_k}$ for $k = 1, \ldots, K$, each consisting of n copies of independent observations from a set of predictor/feature variables of dimension p_k, and also we observe data on q response variables $\mathbf{Y} \in \mathbb{R}^{n \times q}$. Let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_K) \in \mathbb{R}^{n \times p}$ be the design matrix collecting all the predictor sets/groups, with $p = \sum_{k=1}^{K} p_k$. Both p and q can be much larger than the sample size n. Consider the multivariate linear regression model,

$$\mathbf{Y} = \mathbf{X}\mathbf{B}_0 + \mathbf{E} = \sum_{k=1}^{K} \mathbf{X}_k \mathbf{B}_{0k} + \mathbf{E}, \tag{1}$$

where $\mathbf{B}_0 = (\mathbf{B}_{01}^{\mathrm{T}}, \ldots, \mathbf{B}_{0K}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{p \times q}$ is the unknown regression coefficient matrix partitioned corresponding to the predictor groups, and $\mathbf{E}$ contains independent random errors with zero

mean. For simplicity, we assume both the responses and the predictors are centered so there is no intercept term. The naive least squares estimation fails miserably in high dimensions as it leverages neither the response associations nor the grouping of the predictors.

In recent years, we have witnessed an exciting development in regularized estimation, which aims to recover certain parsimonious low dimensional signal from noisy high dimensional data. In the context of multivariate regression or multi-task learning (Caruana, 1997), many exploit the idea of sparse estimation (Rothman et al., 2010; Peng et al., 2010; Lee and Liu, 2012; Li et al., 2015), in which information sharing can be achieved by assuming that all the responses are impacted by the same small subset of predictors. When the predictors themselves exhibit a group structure as in model (1), a group penalization approach, for example, the convex group Lasso (grLasso) method (Yuan and Lin, 2006), can be readily applied to promote groupwise predictor selection. Such methods have shown to be effective in integrative analysis of high-throughput genomic studies (Ma et al., 2011; Liu et al., 2014); a comprehensive review of these methods is provided by Huang et al. (2012).

For multivariate learning, another class of methods, i.e., the reduced-rank methods (Anderson, 1951; Reinsel and Velu, 1998), has also been attractive, where a low-rank constraint on the parameter matrix directly translates to an interpretable latent factor formulation, and conveniently induces information sharing among the regression tasks. Bunea et al. (2011) cast the high-dimensional reduced-rank regression (RRR) as a non-convex penalized regression problem with a rank penalty. Its convex counterpart is the nuclear norm penalized regression (NNP) (Yuan et al., 2007; Negahban and Wainwright, 2011; Koltchinskii et al., 2011),

$$\min_{\mathbf{B}\in\mathbb{R}^{p\times q}} \frac{1}{2n}\|\mathbf{Y} - \mathbf{XB}\|_{\mathbb{F}}^2 + \lambda\|\mathbf{B}\|_{\star}, \tag{2}$$

where $\|\cdot\|_{\mathbb{F}}$ denotes the Frobenius norm, and the nuclear norm is defined as $\|\mathbf{B}\|_{\star} = \sum_{j=1}^{p\wedge q} \sigma(\mathbf{B}, j)$, with $\sigma(\cdot, j)$ denoting the jth largest singular value of the enclosed matrix. Other forms of singular value penalization were considered in, e.g., Mukherjee and Zhu (2011), Chen et al. (2013) and Zhou and Li (2014). In addition, some recent efforts further improve low-rank methods by incorporating error covariance modeling, such as envelope models (Cook et al., 2015), or by utilizing variable selection (Chen et al., 2012; Bunea et al.,

2012; Chen and Huang, 2012; Su et al., 2016).

In essence, to best predict the multivariate response, sparse methods search for the most relevant subset or groups of predictors, while reduced-rank methods search for the most relevant subspace of the predictors. However, neither class of existing methods can fulfill the needs in the aforementioned multi-view problems. The predictors within each group/view may be strongly correlated, each individual variable may only have weak predictive power, and it is likely that only a few of the views are useful for prediction. Indeed, in the lung study, it is largely the collective effort of the sets of local airway features that drives the global lung functions (Chen et al., 2016). In the LSOA study, the predictor groups have distinct interpretations and thus warrant distinct dependence structures with the health outcomes.

In this paper, we propose an *integrative multi-view reduced-rank regression* (iRRR) model, where the integration is in terms of multi-view predictors. To be specific, under model (1), we assume each set of predictors has its own low-rank coefficient matrix. Figure 1 shows a conceptual diagram of our proposed method. Latent features or relevant subspaces are extracted from each predictor set $\mathbf{X}_k$ under the supervision of the multivariate response $\mathbf{Y}$, and the sets of latent variables/subspaces in turn jointly predict $\mathbf{Y}$. The model setting strikes a balance between flexibility and parsimony, as it nicely bridges two seemingly quite different model classes: reduced-rank and group-sparse models. On the one hand, iRRR generalizes the two-set regressor model studied in Velu (1991) by allowing multiple sets of predictors, each of which can correspond to a low-rank coefficient matrix. On the other hand, iRRR subsumes group-sparse model setup by allowing the rank of $\mathbf{B}_{0k}$ being 0, for any $k = 1, \ldots, K$, i.e., the coefficient matrix of a predictor group could be entirely zero.

In Section 2, we develop a new convex optimization approach via composite nuclear norm penalization (cNNP) to conduct model estimation for iRRR, which ensures the scalability to large-scale applications. We devise an Alternating Direction Method of Multipliers (ADMM) algorithm to solve the optimization problem with convergence guarantee; extensions to non-Gaussian response, incomplete data, among others, are also considered, and all the details are reported in Section 3. In Section 4, we derive non-asymptotic oracle bounds for the iRRR estimator, which subsume the results for several existing regularized estimation methods, and show that our proposed approach can achieve superior performance under realistic settings
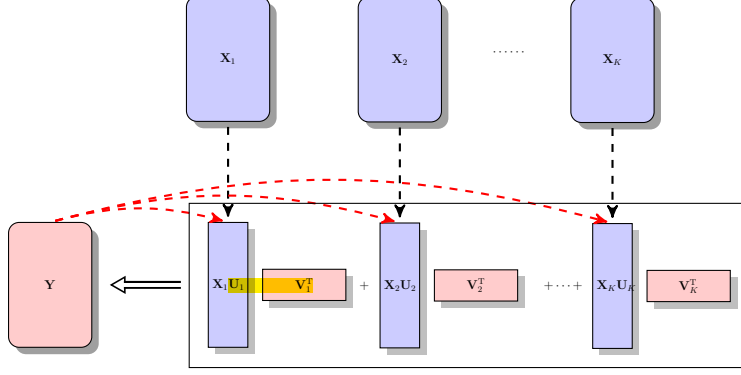
4

Figure 1: A diagram of integrative multi-view reduced-rank regression (iRRR). Latent features, i.e., $\mathbf{X}_k\mathbf{U}_k$, are learned from each view/predictor set under the supervision of $\mathbf{Y}$.

of multi-view learning. Comprehensive simulation studies are contained in Section 5, and a real data analysis of the LSOA example is contained in Section 6. In Section 7, we conclude with some discussions.

2 Integrative Multi-View Reduced-Rank Regression

2.1 Proposed Model

We consider the multivariate regression model in (1) to pursue integrative learning. Recall that in model (1), there are K views or groups of predictors denoted by $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_K)$, where $\mathbf{X}_k \in \mathbb{R}^{n \times p_k}$ and $\sum_{k=1}^{K} p_k = p$. Correspondingly, the coefficient matrix $\mathbf{B}_0$ is partitioned into K parts as $\mathbf{B}_0 = (\mathbf{B}_{01}^{\mathrm{T}}, \ldots, \mathbf{B}_{0K}^{\mathrm{T}})^{\mathrm{T}}$, where $\mathbf{B}_{0k} \in \mathbb{R}^{p_k \times q}$. Denote $r(\cdot)$ as the rank of the enclosed matrix. By assuming each $\mathbf{B}_{0k}$ is possibly of low rank or even a zero matrix, i.e., $0 \le r_{0k} \ll p_k \wedge q$ where $r_{0k} = r(\mathbf{B}_{0k})$, for $k = 1, \ldots, K$, we reach our proposed *integrative multi-view reduced-rank regression* (iRRR) model.

The groupwise low-rank structure in iRRR is distinct from a globally low-rank structure for $\mathbf{B}_0$ in standard RRR models. The low-rankness of $\mathbf{B}_{0k}$s does not necessarily imply that $\mathbf{B}_0$ is of low rank. Conversely, if $\mathbf{B}_0$ is of low rank, i.e., $r_0 = r(\mathbf{B}_0) \ll p \wedge q$, all we know is that the rank of each $\mathbf{B}_{0k}$ is upper bounded by r_0.

Nevertheless, we can first attempt an intuitive understanding of the potential parsimony of iRRR in multi-view settings. The numbers of free parameters in $\mathbf{B}_0$ (the naive degrees of

freedom) for an iRRR model, a globally reduced-rank model and a group-sparse model are $\mathrm{df}_1 = \sum_{k=1}^K (p_k + q - r_{0k})r_{0k}$, $\mathrm{df}_2 = (p + q - r_0)r_0$ and $\mathrm{df}_3 = \sum_{k=1}^K p_k q I(r_{0k} \neq 0)$, respectively, where $I(\cdot)$ is an indicator function. For high-dimensional multi-view data, consider the scenario that only ==a few views/predictor groups impact the prediction in a collective way, i.e., r_{0k}s are mostly zero==, and each nonzero r_{0k} could be much smaller than $(p_k \wedge q)$. Then df_1 could be substantially smaller than both df_2 and df_3. For example, if $r_{01} > 0$ while $r_{0k} = 0$ for any $k > 1$ (i.e., $r_0 = r_{01}$), we have $\mathrm{df}_1 = (p_1 + q - r_{01})r_{01}$, $\mathrm{df}_2 = (p + q - r_{01})r_{01}$ and $\mathrm{df}_3 = p_1 q$, respectively. Another example is when $r_0 = \sum_{k=1}^K r_{0k}$, e.g., $\mathbf{B}_{0k}$s in model (1) have distinct row spaces. Since $\sum_{k=1}^K (p_k + q - r_{0k})r_{0k} \leq \{q + \sum_{k=1}^K (p_k - r_{0k})\}\{\sum_{k=1}^K r_{0k}\} = (p + q - r_0)r_0$, iRRR is more parsimonious than the globally reduced-rank model. The above observations will be rigorously justified in Section 4 through a non-asymptotic analysis.

2.2 Composite Nuclear Norm Penalization

To recover the desired view-specific low-rank structure in the iRRR model, we propose a convex optimization approach with ==*composite nuclear norm penalization*== (cNNP),

$$\widehat{\mathbf{B}} \in \arg \min_{\mathbf{B} \in \mathbb{R}^{p \times q}} \frac{1}{2n} \|\mathbf{Y} - \mathbf{XB}\|_{\mathbb{F}}^2 + \lambda \sum_{k=1}^K w_k \|\mathbf{B}_k\|_\star, \tag{3}$$

where $\|\mathbf{B}_k\|_\star = \sum_{j=1}^{p_k \wedge q} \sigma(\mathbf{B}_k, j)$ is the nuclear norm of $\mathbf{B}_k$, w_ks are some prespecified weights, and λ is a tuning parameter controlling the amount of regularization. The use of the weights is to adjust for the dimension and scale differences of $\mathbf{X}_k$s. We choose

$$w_k = ==\sigma(\mathbf{X}_k, 1)==\{\sqrt{q} + \sqrt{r(\mathbf{X}_k)}\}/n, \tag{4}$$

based on a concentration inequality of the largest singular value of a Gaussian matrix. <u>This choice balances the penalization of different</u> views and allows us to use only a single tuning parameter to achieve desired statistical performance; see Section 4 for details.

Through cNNP, the proposed approach can achieve ==view selection and view-specific subspace selection simultaneously==, which shares the same spirit as the bi-level selection methods for univariate regression (Breheny and Huang, 2009; Huang et al., 2012; Chen et al., 2016). Moreover, iRRR seamlessly bridges group-sparse and low-rank methods as its special cases.

Case 1: *nuclear norm penalized regression (NNP)*. When $p_1 = p$ and $K = 1$, (3) reduces to the NNP method as in (2), which learns a globally low-rank association structure.

Case 2: *multi-task learning (MTL)*. When $p_k = 1$ and $p = K$, (3) becomes a special case of MTL (Caruana, 1997), in which all the tasks are with the same set of features and the same set of samples. MTL achieves integrative learning by exploiting potential information sharing across the tasks, i.e., all the task models share the same sparsity pattern of the features.

Case 3: *Lasso and grLasso*. When $q = 1$, (3) becomes a grLasso method, as $\|\mathbf{B}_k\|_\star = \|\mathbf{B}_k\|_2$ when $\mathbf{B}_k \in \mathbb{R}^{p_k}$. Further, when $p_k = 1$ and $p = K$, (3) reduces to a Lasso regression.

3 Computation and Extensions

3.1 ADMM for iRRR

Without loss of generality, we omit the weights w_k $(k = 1, \cdots, K)$ defined in (4) in the following derivation of the computational algorithm (since we can reparameterize $\mathbf{X}_k$ by $(1/w_k)\mathbf{X}_k$ and $w_k \mathbf{B}_k$ by $\mathbf{B}_k$ to get an equivalent unweighted form of the objective function). The convex optimization has no closed-form solution, for which we propose an ADMM algorithm (Boyd et al., 2011). More specifically, let $\mathbf{A}_k$ $(k = 1, \cdots, K)$ be a set of surrogate variables for $\mathbf{B}_k$ with the same dimensions and $\mathbf{A} = (\mathbf{A}_1^{\mathrm{T}}, \cdots, \mathbf{A}_K^{\mathrm{T}})^{\mathrm{T}}$. The original optimization is equivalent to

$$\min_{\mathbf{A}_k, \mathbf{B}_k} \quad \frac{1}{2n}\|\mathbf{Y} - \sum_{k=1}^{K} \mathbf{X}_k \mathbf{B}_k\|_{\mathbb{F}}^2 + \lambda \sum_{k=1}^{K} \|\mathbf{A}_k\|_\star, \qquad s.t. \quad \mathbf{A}_k = \mathbf{B}_k, \quad k = 1, \cdots, K.$$

Let $\mathbf{\Lambda}_k$ $(k = 1, \cdots, K)$ be a set of Lagrange multipliers with the same dimensions as $\mathbf{A}_k$ and $\mathbf{B}_k$, and $\mathbf{\Lambda} = (\mathbf{\Lambda}_1^{\mathrm{T}}, \cdots, \mathbf{\Lambda}_K^{\mathrm{T}})^{\mathrm{T}}$. The augmented Lagrangian objective function is

$$\begin{aligned}
\mathcal{D}(\mathbf{Y}; \mathbf{A}, \mathbf{B}, \mathbf{\Lambda}) =& \frac{1}{2n}\|\mathbf{Y} - \sum_{k=1}^{K} \mathbf{X}_k \mathbf{B}_k\|_{\mathbb{F}}^2 + \lambda \sum_{k=1}^{K} \|\mathbf{A}_k\|_\star \\
&+ \sum_{k=1}^{K} \langle \mathbf{\Lambda}_k, \mathbf{A}_k - \mathbf{B}_k \rangle_{\mathbb{F}} + \frac{\rho}{2} \sum_{k=1}^{K} \|\mathbf{A}_k - \mathbf{B}_k\|_{\mathbb{F}}^2,
\end{aligned} \tag{5}$$

where $\langle \mathbf{Q}, \mathbf{R} \rangle_{\mathbb{F}}$ represents the Frobenius inner product of $\mathbf{Q}$ and $\mathbf{R}$, which equals to the trace of $\mathbf{Q}^{\mathrm{T}}\mathbf{R}$. The last squared Frobenius term is the augmentation term, with ρ being a

prespecified step size (usually set to be a small positive value, e.g., 0.1).

The ADMM algorithm alternates between two steps, a *primal* step and a *dual* step, until convergence. The primal step minimizes $\mathcal{D}(\mathbf{Y}; \mathbf{A}, \mathbf{B}, \boldsymbol{\Lambda})$ with respect to $\mathbf{A}$ and $\mathbf{B}$, respectively, while fixing everything else; the *dual* step updates $\boldsymbol{\Lambda}$.

Primal step: We minimize (5) with respect to $\mathbf{A}$ and $\mathbf{B}$, separately. In particular, when one is fixed, the optimization with respect to the other has an explicit solution. More specifically, let $\widetilde{\mathbf{A}}$, $\widetilde{\mathbf{B}}$, and $\widetilde{\boldsymbol{\Lambda}}$ represent the estimates from the previous iteration. The optimization $\min_{\mathbf{B}} \mathcal{D}(\mathbf{Y}; \widetilde{\mathbf{A}}, \mathbf{B}, \widetilde{\boldsymbol{\Lambda}})$ has a unique solution

$$\widehat{\mathbf{B}} = \left(\frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X} + \rho\mathbf{I}\right)^{-1}\left(\frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{Y} + \rho\widetilde{\mathbf{A}} + \widetilde{\boldsymbol{\Lambda}}\right). \tag{6}$$

Subsequently, we can obtain the estimate of $\mathbf{B}_k$ (i.e., $\widehat{\mathbf{B}}_k$) by partitioning $\widehat{\mathbf{B}}$.

To estimate $\mathbf{A}$, the objective function $\mathcal{D}(\mathbf{Y}; \mathbf{A}, \widehat{\mathbf{B}}; \widetilde{\boldsymbol{\Lambda}})$ is readily separable for different $\mathbf{A}_k$s. In particular, each subproblem is rewritten as

$$\min_{\mathbf{A}_k} \quad \frac{\rho}{2}\|\mathbf{A}_k - \widehat{\mathbf{B}}_k + \frac{\widetilde{\boldsymbol{\Lambda}}_k}{\rho}\|_{\mathbb{F}}^2 + \lambda\|\mathbf{A}_k\|_{\star}, \tag{7}$$

which can be solved via the singular value ==soft-thresholding technique== (Cai et al., 2010). To be specific, let $\mathbf{U}_k\mathbf{D}_k\mathbf{V}_k^{\mathrm{T}}$ be the singular value decomposition (SVD) of $\widehat{\mathbf{B}}_k - \widetilde{\boldsymbol{\Lambda}}_k/\rho$, where $\mathbf{U}_k$ and $\mathbf{V}_k$ have orthonormal columns and $\mathbf{D}_k$ contains non-increasing singular values. The solution to the optimization problem in (7) is

$$\widehat{\mathbf{A}}_k = \mathbf{U}_k\mathcal{S}(\mathbf{D}_k, \frac{\lambda}{\rho})\mathbf{V}_k^{\mathrm{T}}, \tag{8}$$

where $\mathcal{S}(\mathbf{D}_k, \lambda/\rho) = (\mathbf{D}_k - \lambda/\rho)_+$ applies soft-thresholding at the level λ/ρ to each entry of $\mathbf{D}_k$. As a result, $\widehat{\mathbf{A}}_k$ may be low-rank.

Dual step: Once $\mathbf{A}$ and $\mathbf{B}$ are estimated, the Lagrange multipliers $\boldsymbol{\Lambda}_k$ are updated by

$$\widehat{\boldsymbol{\Lambda}}_k = \widetilde{\boldsymbol{\Lambda}}_k + \rho(\widehat{\mathbf{A}}_k - \widehat{\mathbf{B}}_k). \tag{9}$$

Stopping criterion: The ADMM algorithm alternates between the primal step and the dual step. After each iteration, we evaluate the primal and dual residuals as

$$r_{primal} = \|\widehat{\mathbf{A}} - \widehat{\mathbf{B}}\|_{\mathbb{F}}, \quad r_{dual} = \rho\|\widehat{\mathbf{B}} - \widetilde{\mathbf{B}}\|_{\mathbb{F}}. \tag{10}$$

Following Boyd et al. (2011), the stopping criterion is that both residuals fall below a small prefixed threshold. It can be proved that under weak regularity conditions, the algorithm always converges to a global optimum. In practice, one can let the step size ρ vary over iterations, and generally the convergence is expedited with a slowly increasing sequence of ρ (He et al., 2000). A summary of the above algorithm for solving iRRR with a fixed λ is provided in Algorithm 1 below.

Algorithm 1 ADMM algorithm for fitting iRRR

Parameter: λ, ρ.

Initialize $\mathbf{A}$, $\mathbf{B}$ and the Lagrange multiplier $\mathbf{\Lambda}$;

while The stopping criterion is not satisfied **do**

- Primal step: update $\mathbf{B}$ by (6) and update $\mathbf{A}$ by (8);

- Dual step: update $\mathbf{\Lambda}$ by (9);

- Calculate the primal and dual residuals in (10);

- (Optional) Increase ρ by a small amount, e.g., $\rho \leftarrow 1.1\rho$.

end while

The tuning parameter λ in (3) balances the loss function and the penalty term. In practice, the model is fitted using the ADMM algorithm for a sequence of λ values to produce a spectrum of view-specific low-rank models. A warm start strategy is adopted to speed up computation, i.e., the current solution is used as the initial value for the next λ value. We use K-fold cross validation (Stone, 1974) to choose the optimal λ and hence the optimal solution, based on the predictive performance of the models.

3.2 Handling Non-Gaussian and Incomplete Response

When the responses are non-Gaussian, we substitute the squared loss function in (3) with the negative log likelihood denoted as $-\log L(\mathbf{Y}, \mathbf{\Theta})$. The augmented Lagrangian becomes

$$\mathcal{D}(\mathbf{Y}; \boldsymbol{\mu}, \mathbf{A}, \mathbf{B}, \mathbf{\Lambda}) = -\frac{1}{n}\log L(\mathbf{Y}, \mathbf{\Theta}) + \lambda\sum_{k=1}^{K}\|\mathbf{A}_k\|_\star + \langle\mathbf{\Lambda}, \mathbf{A} - \mathbf{B}\rangle_\mathbb{F} + \frac{\rho}{2}\|\mathbf{A} - \mathbf{B}\|_\mathbb{F}^2,$$

where $\boldsymbol{\Theta} = \mathbf{1}\boldsymbol{\mu}^{\mathrm{T}} + \mathbf{X}\mathbf{B}$. The minimization of $\mathcal{D}(\mathbf{Y}; \boldsymbol{\mu}, \mathbf{A}, \mathbf{B}, \boldsymbol{\Lambda})$ with respect to $\boldsymbol{\mu}$ and $\mathbf{B}$ while fixing everything else may no longer have closed-form solutions. To alleviate the computational burden, one could apply a quadratic approximation or majorization to the negative log likelihood function in the primal step, and then follow the ADMM algorithm for parameter estimation. In the following, we demonstrate the estimation procedure for binary responses.

The log-likelihood function for binary responses $\mathbf{Y}$ can be expressed as

$$\log L(\mathbf{Y}, \boldsymbol{\Theta}) = \sum_{i=1}^{n} \sum_{j=1}^{q} \log h\left((2y_{ij} - 1)\theta_{ij}\right), \tag{11}$$

where θ_{ij} is the (i,j)th entry of $\boldsymbol{\Theta}$ and $h(\eta) = \exp(\eta)/\{1 + \exp(\eta)\}$ denotes the inverse function of the logit link function. Following Lee et al. (2010) and Lee and Huang (2013), we have the following relation

$$-\log h(\eta) \leq -\log h(\eta_0) - 2\{1 - h(\eta_0)\}^2 + \frac{1}{8}\left[\eta - \eta_0 - 4\{1 - h(\eta_0)\}\right]^2. \tag{12}$$

Namely, $-\log h(\eta)$ is majorized by the quadratic function on the right-hand side, which is tangent with $-\log h(\eta)$ at η_0 and has a fixed second-order derivative.

Let $\widetilde{\theta}_{ij}$ be the estimate from the previous iteration. By applying (12) to (11), we have

$$-\log L(\mathbf{Y}, \boldsymbol{\Theta}) \leq \frac{1}{8} \sum_{i=1}^{n} \sum_{j=1}^{q} \left[(2y_{ij} - 1)(\theta_{ij} - \widetilde{\theta}_{ij}) - 4\left\{1 - h\left((2y_{ij} - 1)\widetilde{\theta}_{ij}\right)\right\}\right]^2 + c,$$

where c is some constant. Let $\mathbf{Y}^{\star}$ be an $n \times q$ working response matrix with the (i,j)th entry

$$y_{ij}^{\star} = \widetilde{\theta}_{ij} + 4(2y_{ij} - 1)\left\{1 - h\left((2y_{ij} - 1)\widetilde{\theta}_{ij}\right)\right\}.$$

Correspondingly, the negative log likelihood function $-\log L(\mathbf{Y}, \boldsymbol{\Theta})$ is majorized by the squared function $1/8\|\mathbf{Y}^{\star} - \boldsymbol{\Theta}\|_{\mathbb{F}}^2$, plus some constant. Consequently, in the primal step, one could minimize the majorized objective function to estimate $\boldsymbol{\mu}$ and $\mathbf{B}$ explicitly. In particular, the estimate of $\boldsymbol{\mu}$ is $(1/n)\mathbf{Y}^{\star\mathrm{T}}\mathbf{1}$. We remark that in practice, it generally suffices to run the majorization-minimization procedure once in each ADMM iteration (He et al., 2002).

When there are missing values in the responses, we exploit a similar idea to majorize the objective function in each ADMM iteration. More specifically, suppose $\mathcal{O} \subseteq \{(i,j) :$

$i = 1, \cdots, n; j = 1, \cdots, q\}$ is the index set for observed data points, and $\mathcal{M} \subseteq \{(i,j) : i = 1, \cdots, n; j = 1, \cdots, q\}$ is the index set for missing values. For Gaussian data, we majorize the observed loss function $\sum_{(i,j)\in\mathcal{O}}(y_{ij} - \theta_{ij})^2$ by $\sum_{(i,j)\in\mathcal{O}}(y_{ij} - \theta_{ij})^2 + \sum_{(i,j)\in\mathcal{M}}(\widetilde{\theta}_{ij} - \theta_{ij})^2$; for binary responses, we first majorize the negative log likelihood function by $1/8\sum_{(i,j)\in\mathcal{O}}(y_{ij}^{\star} - \theta_{ij})^2$ as before, and then further majorize it as $1/8\sum_{(i,j)\in\mathcal{O}}(y_{ij}^{\star} - \theta_{ij})^2 + 1/8\sum_{(i,j)\in\mathcal{M}}(\widetilde{\theta}_{ij} - \theta_{ij})^2$. By collecting y_{ij} or $y_{ij}^{\star}$ and $\widetilde{\theta}_{ij}$ in an $n \times p$ matrix, we obtain a matrix-form loss function as before. As a result, we use the same ADMM steps to estimate the parameters.

3.3   On ℓ_2 Regularization and Adaptive Estimation

To better deal with high dimensional data, we can consider adding a ridge penalty $\lambda_2\|\mathbf{B}\|_{\mathbb{F}}^2$ to the cNNP penalty in (3) (Mukherjee and Zhu, 2011; Chen et al., 2013). As a result, the objective function becomes strictly convex whenever the tuning parameter $\lambda_2 > 0$. This shares the same idea as the elastic net (Zou and Hastie, 2005), and ensures that the problem has a unique global optimizer.

With the combined penalty form $\lambda \sum_{k=1}^{K} \|\mathbf{B}_k\|_{\star} + \lambda_2\|\mathbf{B}\|_{\mathbb{F}}^2$, the iRRR problem can be easily transformed to the same form as before:

$$\frac{1}{2n}\left\|\begin{pmatrix}\mathbf{Y}\\\mathbf{0}\end{pmatrix} - \begin{pmatrix}\mathbf{X}\\\sqrt{2n\lambda_2}\mathbf{I}\end{pmatrix}\mathbf{B}\right\|_{\mathbb{F}}^2 + \lambda\sum_{k=1}^{K}\|\mathbf{B}_k\|_{\star},$$

where $\mathbf{0}$ is a zero matrix of size $p \times q$ and $\mathbf{I}$ is an identity matrix of size $p \times p$. (More generally the identity matrix can be replaced by a diagonal matrix to allow weighted ℓ_2 regularization). Upon defining $\mathbf{Y}^{\dagger} = (\mathbf{Y}^{\mathrm{T}}, \mathbf{0})^{\mathrm{T}}$ and $\mathbf{X}^{\dagger} = (\mathbf{X}^{\mathrm{T}}, \sqrt{2n\lambda_2}\mathbf{I})^{\mathrm{T}}$ as augmented responses and predictors, the model estimation could be conducted directly by applying Algorithm 1 to the augmented data. Alternatively, a more computationally efficient way is to directly modify the ADMM algorithm by replacing the nuclear norm penalty in (7) by a combined nuclear and squared ℓ_2 norm penalty. The resulting problem can still be solved explicitly, now via a singular value shrinkage and thresholding operation (Sun and Zhang, 2012).

When the ridge penalty is included, we have an additional tuning parameter λ_2. A larger value of λ_2 makes the problem more convex, but meanwhile introduces more bias to

the final estimates. In practice, λ_2 can be selected using CV as well. However, empirical experiments suggest that it usually suffices to set λ_2 at a very small value without tuning it. For simplicity, we omit the ridge penalty term in our numerical studies.

Moreover, motivated by Zou (2006), we can consider an adaptively weighted version of iRRR, where, for example, we first fit iRRR and then adjust the weights according to the estimated coefficient sub-matrices (e.g., factoring in the inverse of the Frobenius norms of the estimated coefficient matrices). This may potential improve view selection and predictive accuracy, as shown in the numerical studies in Section 5.

4   Theoretical Analysis

We investigate the theoretical properties of the proposed iRRR estimator from solving the convex cNNP problem. In particular, we derive its non-asymptotic performance bounds for estimation and prediction. Our general results recover performance bounds of several related methods, including Lasso, grLasso and NNP. We further show that iRRR is capable of substantially outperforming those methods under realistic settings of multi-view learning. All the proofs are provided in Appendix A.

We mainly consider the multi-view regression model in (1), i.e., $\mathbf{Y} = \sum_{k=1}^{K} \mathbf{X}_k \mathbf{B}_{0k} + \mathbf{E}$, and the iRRR estimator in (3) with the weights defined in (4), i.e.,

$$\widehat{\mathbf{B}} \in \arg\min_{\mathbf{B} \in \mathbb{R}^{p \times q}} \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_{\mathbb{F}}^2 + \lambda \sum_{k=1}^{K} \sigma(\mathbf{X}_k, 1) \left\{ \sqrt{q} + \sqrt{r(\mathbf{X}_k)} \right\} \|\mathbf{B}_k\|_{\star}/n.$$

Define $\mathbf{Z} = \mathbf{X}^{\mathrm{T}} \mathbf{X}/n$, and $\mathbf{Z}_k = \mathbf{X}_k^{\mathrm{T}} \mathbf{X}_k/n$, for $k = 1, \dots, K$. We scale the columns of $\mathbf{X}$ such that the diagonal elements of $\mathbf{Z}$ all equal to 1. Denote $\Lambda(\mathbf{Z}, l)$ as the lth largest eigenvalue of $\mathbf{Z}$, so that $\Lambda(\mathbf{Z}, l) = \sigma(\mathbf{X}, l)^2/n$.

Theorem 4.1. *Assume $\mathbf{E}$ has independent and identically distributed (i.i.d.) $N(0, \tau^2)$ entries. Let $\lambda = (1+\theta)\tau$, with $\theta > 0$ arbitrary. Then with probability at least $1 - \sum_{k=1}^{K} \exp[-\theta^2\{q + r(\mathbf{X}_k)\}/2]$, we have*

$$\|\mathbf{X}\widehat{\mathbf{B}} - \mathbf{X}\mathbf{B}_0\|_{\mathbb{F}}^2 \leq \|\mathbf{X}\mathbf{C} - \mathbf{X}\mathbf{B}_0\|_{\mathbb{F}}^2 + 4\lambda \sum_{k=1}^{K} \sigma(\mathbf{X}_k, 1) \left\{ \sqrt{q} + \sqrt{r(\mathbf{X}_k)} \right\} \|\mathbf{C}_k\|_{\star},$$

for any $\mathbf{C}_k \in \mathbb{R}^{p_k \times q}$, $k = 1, \dots, K$ and $\mathbf{C} = (\mathbf{C}_1^{\mathrm{T}}, \dots, \mathbf{C}_K^{\mathrm{T}})^{\mathrm{T}}$.

12

Theorem 4.1 shows that $\widehat{\mathbf{B}}$ balances the bias term $\|\mathbf{XC} - \mathbf{XB}_0\|_{\mathbb{F}}^2$ and the variance term $4\lambda \sum_{k=1}^{K} \sigma(\mathbf{X}_k, 1)\{\sqrt{q} + \sqrt{r(\mathbf{X}_k)}\}\|\mathbf{C}_k\|_{\star}$. An oracle inequality for $\widehat{\mathbf{B}}$ is then readily obtained for the low-dimensional scenario $\sigma(\mathbf{X}, p) > 0$; see the corollary in Appendix A.3.

We now investigate the general high-dimensional scenario. Motivated by Lounici et al. (2011), Negahban and Wainwright (2011), Koltchinskii et al. (2011), among others, we impose a restricted eigenvalue condition (RE). We say that $\mathbf{X}$ satisfies RE condition over a restricted set $\mathcal{C}(r_1, \ldots, r_K; \delta) \subset \mathbb{R}^{p \times q}$ if there exists some constant $\kappa(\mathbf{X}) > 0$ such that

$$\frac{1}{2n}\|\mathbf{X}\Delta\|_{\mathbb{F}}^2 \geq \kappa(\mathbf{X})\|\Delta\|_{\mathbb{F}}^2, \qquad \text{for all } \Delta \in \mathcal{C}(r_1, \ldots, r_K; \delta).$$

Here each r_k is an integer satisfying $1 \leq r_k \leq \min(p_k, q)$ and δ is a tolerance parameter. The technical details on the construction of the restricted set is provided in Appendix A.1.

Theorem 4.2. *Assume that $\mathbf{E}$ has i.i.d. $N(0, \tau^2)$ entries. Suppose $\mathbf{X}$ satisfies the RE condition with parameter $\kappa(\mathbf{X}) > 0$ over the set $\mathcal{C}(r_1, \ldots, r_K; \delta)$. Let $\lambda = 2(1 + \theta)\tau$ with $\theta > 0$ arbitrary. Then with probability at least $1 - \sum_{k=1}^{K} \exp[-\theta^2\{q + r(\mathbf{X}_k)\}/2]$,*

$$\|\widehat{\mathbf{B}} - \mathbf{B}_0\|_{\mathbb{F}}^2 \preceq \max\left\{\delta^2, \tau^2(1+\theta)^2\sum_{k=1}^{K}\frac{\Lambda(\mathbf{Z}_k, 1)}{\kappa(\mathbf{X})^2}\frac{\{\sqrt{q} + \sqrt{r(\mathbf{X}_k)}\}^2 r_k}{n},\right.$$
$$\left.\tau(1+\theta)\sum_{k=1}^{K}\frac{\sqrt{\Lambda(\mathbf{Z}_k, 1)}}{\kappa(\mathbf{X})}\frac{\{\sqrt{q} + \sqrt{r(\mathbf{X}_k)}\}\{\sum_{j=r_k+1}^{m_k}\sigma(\mathbf{B}_{0k}, j)\}}{\sqrt{n}}\right\}.$$

On the right hand side of the above upper bound, the first term is from the tolerance parameter in the RE condition, which ensures that the condition can possibly hold when the true model is not exactly low-rank (Negahban and Wainwright, 2011), i.e., when $\sum_{j=r_k+1}^{m_k}\sigma(\mathbf{B}_{0k}, j) \neq 0$. The second term gives the *estimation error* of recovering the desired view-specific low-rank structure, and the third term gives the *approximation error* incurred due to approximating the true model with the view-specific low-rank structure. When the true model is exactly of low rank, i.e., $r(B_{0k}) = r_{0k}$, it suffices to take $\delta = 0$ and the upper bound then yields the estimation error, i.e., $\tau^2\sum_{k=1}^{K}\{q + r(\mathbf{X}_k)\}r_{0k}/n$. This rate holds with high probability in the high-dimensional setting that $q + r(\mathbf{X}_k) \to \infty$. In the classical setting of $n \to \infty$ with fixed q and $r(\mathbf{X}_k)$, by choosing $\theta \propto \sqrt{\log n}$, the rate becomes $\tau^2 \log(n) \sum_{k=1}^{K} r_{0k}/n$ with probability approaching 1.

Intriguingly, the results in Theorem 4.2 can specialize into oracle inequalities of several existing regularized estimation methods, such as NNP, MTL and Lasso. This is because these models can all be viewed as special cases of iRRR. As such, iRRR seamlessly bridges group-sparse and low-rank methods and provides a unified theory of the two types of regularization. Several examples are provided in Appendix A.2.

To see the potential advantage of iRRR over NNP or MTL, we make some comparisons of their error rates based on Theorem 4.2. To convey the main message, consider the case where $p_k = p_1$, $r(\mathbf{X}_k) = r_{X_1}$ for $k = 1, \ldots, K$, $r_{0k} = r_{01}$ for $k = 1, \ldots, s$, and $r_{0k} = 0$ for $k = s + 1, \ldots, K$. The error rate is $\tau^2 s r_{01}(q + r_{X_1})/n$, $\tau^2 r_0(q + r_X)/n$, for iRRR and NNP, respectively, with high probability. As long as $s r_{01} = O(r_0)$, iRRR achieves a faster rate since $r_{X_1} \le r_X$ always holds. For comparing iRRR and MTL, we get that with probability $1 - p^{-1}$, iRRR achieves an error rate $\tau^2(\log p + q + r_{X_1}) s r_{01}/n$ (by choosing $\theta = \sqrt{4 \log p/(q + r_{X_1})}$) while MTL achieves $\tau^2(\log p + q + 1) s p_1/n$. The two rates agree with each other in the MTL setting when $r_{X_1} = r_{01} = p_1 = 1$, and the former rate can be much faster in the iRRR setting when, for example, $r_{01} \ll p_1$ and $r_{X_1} = O(\log(p) + q)$.

5   Simulation

5.1   Settings and Evaluation Metrics

We conduct simulation studies to demonstrate the efficacy of the proposed iRRR method. We consider two response types: Gaussian and binary. In Gaussian settings, we compare iRRR with the ordinary least squares (OLS), the ridge RRR (RRRR) (Mukherjee and Zhu, 2011) (which contains RRR as a special case), and the adaptive NNP (aNNP) (which has been shown to be computationally efficient and can outperform NNP in Chen et al., 2013). For the settings in which the true coefficient matrix is sparse, we also include MTL (Caruana, 1997) (by treating each predictor as a group in iRRR), as well as Lasso (Tibshirani, 1996) and grLasso (Yuan and Lin, 2006) for each response variable separately (grLasso accounts for the grouping information in the multi-view predictors). In binary settings, we compare iRRR with the generalized RRR (gRRR) (She, 2013; Luo et al., 2018) and the univariate

penalized logistic regression (glmnet) with the <mark>elastic net penalty</mark> (Zou and Hastie, 2005).

For the Gaussian models, we consider a range of simulation settings. **Setting 1** is the basic setting, where $n = 500$, $K = 2$, $p_1 = p_2 = 50$ ($p = 100$), and $q = 100$. We generate the rows of the design matrix $\mathbf{X}$ independently from a p-variate Gaussian distribution $\mathrm{N}(\mathbf{0}, \mathbf{\Sigma}_x)$ with $\mathbf{\Sigma}_x = \mathbf{I}_p$, followed by column centering. The error matrix $\mathbf{E}$ is filled with i.i.d. standard Gaussian random numbers. (We also consider correlated errors. The results are similar and contained in Appendix B.) Each coefficient matrix $\mathbf{B}_{0k}$ has rank $r_{0k} = 10$, which is generated as $\mathbf{B}_{0k} = \mathbf{L}_k \mathbf{R}_k^{\mathrm{T}}$ with the entries of $\mathbf{L}_k \in \mathbb{R}^{p_k \times r_{0k}}$ and $\mathbf{R}_k \in \mathbb{R}^{q \times r_{0k}}$ both generated from $\mathrm{N}(0,1)$. Consequently, $\mathbf{B}_0 = (\mathbf{B}_{01}^{\mathrm{T}}, \mathbf{B}_{02}^{\mathrm{T}})^{\mathrm{T}}$ has rank $r_0 = r_{01} + r_{02} = 20$. The response matrix $\mathbf{Y}$ is then generated based on the model in (1). As such, there are more than 10,000 unknown parameters in this model, posing a challenging large-scale problem. Furthermore, we also consider incomplete responses, with 10%, 20%, 30% entries missing completely at random.

The other settings are variants of **Setting 1**:

- **Setting 2 (multi-collinear):** The predictors in the two views $\mathbf{X}_1$ and $\mathbf{X}_2$ are highly correlated. All the $p = p_1 + p_2$ predictors are generated jointly from a p-variate Gaussian distribution $\mathrm{N}_p(\mathbf{0}, \mathbf{\Sigma}_x)$, where $\mathbf{\Sigma}_x$ has diagonal elements 1 and off-diagonal 0.9.

- **Setting 3 (globally low-rank):** We set $\mathbf{R}_1 = \mathbf{R}_2$ when generating $\mathbf{B}_{01}$ and $\mathbf{B}_{02}$, so that the low rank structures in separate coefficient matrices also imply a globally low-rank structure. We consider three scenarios: $r_0 = r_{01} = r_{02} = 20$, $r_0 = r_{01} = r_{02} = 40$, and $r_0 = 60, r_{01} = r_{02} = 50$.

- **Setting 4 (multi-set):** We consider multiple views, $K \in \{3, 4, 5\}$. The additional design matrices and coefficient matrices are generated in the same way as in **Setting 1**.

- **Setting 5 (sparse-view):** We consider $K = 3$, where the last predictor set $\mathbf{X}_3$ is generated in the same way as in **Setting 1** but is irrelevant to prediction, i.e., $\mathbf{B}_{03} = 0$.

For the binary models, we consider two settings: the basic setting (**Setting 6**) and the sparse-view setting (**Setting 7**), which are similar to **Setting 1** and **Setting 5**, respectively.

The differences are that the sample size is set to $n = 200$, the intercept $\boldsymbol{\mu}_0$ is set as a vector of random numbers from the uniform distribution on $[-1, 1]$, and the entries of $\mathbf{Y}$ are drawn from Bernoulli distributions with their natural parameters given by $\boldsymbol{\Theta} = \mathbf{1}\boldsymbol{\mu}_0^{\mathrm{T}} + \sum_{k=1}^{K} \mathbf{X}_k \mathbf{B}_{0k}$.

In **Settings 1–5**, we use the MSPE to evaluate the performance of different methods,

$$\mathrm{MSPE}(\mathbf{B}_0, \widehat{\mathbf{B}}) = \mathrm{tr}\left\{ (\mathbf{B}_0 - \widehat{\mathbf{B}})^{\mathrm{T}} \boldsymbol{\Sigma}_x (\mathbf{B}_0 - \widehat{\mathbf{B}}) \right\},$$

where $\mathrm{tr}(\cdot)$ represents the trace of a matrix, $\widehat{\mathbf{B}}$ is the estimate of $\mathbf{B}_0$, and $\boldsymbol{\Sigma}_x$ is the covariance matrix of $\mathbf{X}$. In **Settings 6–7**, we evaluate the average cross entropy between the true and estimated probabilities on an independently generated validation data set of size $n = 500$,

$$\mathrm{En}(\boldsymbol{\mu}_0, \mathbf{B}_0, \widehat{\boldsymbol{\mu}}, \widehat{\mathbf{B}}) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{q} \left\{ p_{ij} \log \widehat{p}_{ij} + (1 - p_{ij}) \log(1 - \widehat{p}_{ij}) \right\},$$

where $p_{ij} = \exp(\theta_{ij}) / \{1 + \exp(\theta_{ij})\}$, and $\widehat{p}_{ij}$ is its corresponding estimate.

For each simulation setting, we first generate an independent testing data set to select tuning parameters for different methods. Once selected, the tuning parameters are fixed in subsequent analyses. This unified approach alleviates inaccuracy in the empirical tuning parameter selection to ensure a fair comparison of different regularization methods. We have also tried 5-fold CV. The results are similar to those from the validation data tuning and thus omitted for brevity. In each setting, the experiment is replicated 100 times.

5.2   Results

Table 1 reports the results for **Settings 1–4**. In all the settings, the three regularized estimation methods always substantially outperform OLS, indicating the strength and necessity of dimension reduction. In **Setting 1 (basic)**, iRRR provides the best prediction performance, followed by aNNP and RRRR. When the outcomes are incomplete, only iRRR is applicable. The mean and standard deviation of MSPE over 100 repetitions are 7.87 (0.20), 8.64 (0.20), and 9.96 (0.24), when 10%, 20%, and 30% of the responses are missing, respectively. In **Setting 2 (multi-collinear)**, iRRR is still the best. It is worth noting that owing to shrinkage estimation, RRRR slightly outperforms aNNP. In **Setting 3 (globally low-rank)**, aNNP and RRRR can slightly outperform iRRR when r_0 is much smaller than $\sum_{k=1}^{K} r_{0k}$. This can be explained by the fact that under this setting iRRR may be less parsimonious than the

16

globally reduced-rank methods. To see this, when r_0 is small and $r_0 = r_{01} = r_{02}$, we have that $\sum_{k=1}^{K}(p_k + q - r_{0k})r_{0k} = \{p + K(q - r_0)\}r_0 > (p + q - r_0)r_0$, i.e., iRRR yields a larger number of free parameters than RRR. Nevertheless, iRRR regains its superiority over the globally low-rank methods when r_0 becomes large. We remark that in multi-view problems the scenario of $r_0 \ll \sum_k r_{0k}$ rarely happens unless the relevant subspace from each view largely overlaps with each other. In **Setting 4 (multi-set)**, we confirm that the advantage of iRRR becomes more obvious as the number of distinct view sets increases.

Table 1: Simulation results for **Settings 1–4**. The mean and standard deviation (in parenthesis) of MSPE over 100 simulation runs are presented. In each setting, the best results are highlighted in boldface.

		iRRR	aNNP	RRRR	OLS
Setting 1		**7.22** (0.17)	7.76 (0.22)	8.38 (0.24)	25.15 (0.36)
Setting 2		**4.21** (0.10)	4.69 (0.11)	4.52 (0.11)	25.15 (0.36)
	$(r_0 = 20)$	10.13 (0.22)	**7.81** (0.25)	8.25 (0.26)	25.16 (0.39)
Setting 3	$(r_0 = 40)$	12.48 (0.19)	**12.39** (0.22)	13.76 (0.26)	25.04 (0.37)
	$(r_0 = 60)$	**13.62** (0.21)	14.66 (0.26)	15.66 (0.17)	25.11 (0.39)
	$(K = 3)$	**10.19** (0.21)	13.99 (0.32)	15.44 (0.31)	43.76 (0.59)
Setting 4	$(K = 4)$	**13.04** (0.22)	19.99 (0.35)	19.68 (0.19)	68.00 (0.89)
	$(K = 5)$	**14.84** (0.25)	24.90 (0.32)	21.43 (0.21)	101.87 (1.38)

Figure 2 displays the results for **Setting 5 (sparse-view)**. We find that the iRRR solution tuned based on predictive accuracy usually estimates the third coefficient matrix (which is a zero matrix in truth) as a nearly zero matrix and occasionally an exact zero matrix; in view of the construction of the cNNN penalty in iRRR, this "over-selection" property is analogous to that of Lasso or grLasso. Motivated by Zou (2006), we also experiment with an adaptive iRRR (denoted by iRRR-a) approach, where we first fit iRRR and then adjust the predefined weights by the inverse of the Frobenius norms of the estimated coefficient matrices. As a result, the iRRR-a approach achieves much improved view selection performance and even better prediction accuracy than iRRR. In contrast, MTL, Lasso and grLasso have

worse performance than the low-rank methods, because they fail to leverage information from the multivariate response and/or multi-view predictor structures.
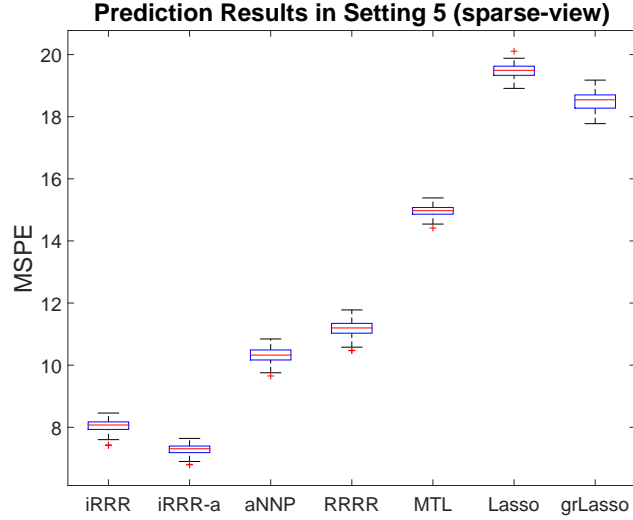

Figure 2: Simulation results for **Setting 5 (sparse-view)**. OLS is omitted as its performance is much worse than the reported methods.

The simulation results of **Settings 6-7** for binary models are displayed in Figure 3. The results are similar as in the Gaussian models, i.e., the iRRR methods substantially outperform the competing sparse or low-rank methods in prediction.
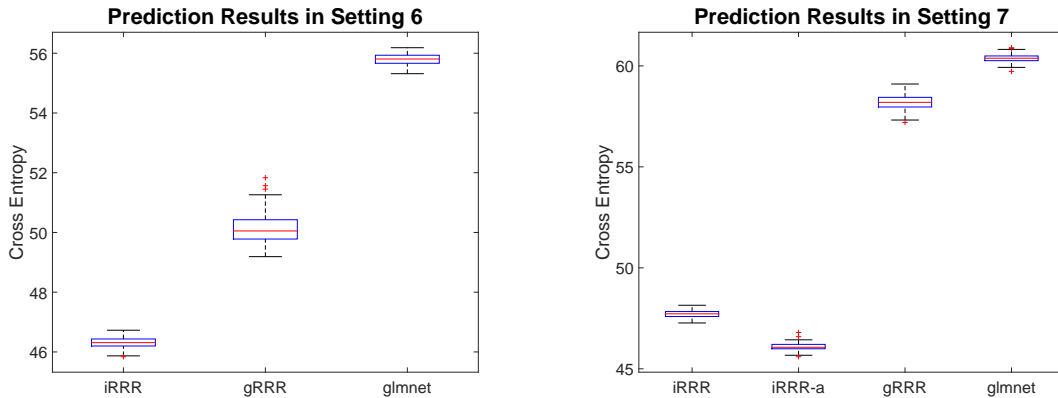
Figure 3: Simulation results for **Settings 6–7** with binary response variables.

We have also compared the computational time of different methods (on a standard

desktop with Intel i5 3.3GHz CPU). For example, the average time (in seconds) under **Setting 1** is 0.68 (0.06), 0.07 (0.01) and 0.02 (0.00) for iRRR, aNNP and RRRR, respectively; under **Setting 4** with $K = 5$ the average time becomes 0.96 (0.12), 0.09 (0.01) and 0.05 (0.01); under **Setting 6** with binary responses, the average time is 1.71 (0.03), 0.98 (0.08) and 0.70 (0.08) for iRRR, gRRR and glmnet. As expected, iRRR is more computationally expensive than the globally low-rank or sparse methods. However, in view of the scale of the problem, the computational cost for iRRR is still low and acceptable. The Matlab code for implementing the proposed method is available at https://github.com/reagan0323/iRRR.

6 An Application in the Longitudinal Studies of Aging

The LSOA (Stanziano et al., 2010) was a collaborative effort of the National Center for Health Statistics and the National Institute on Aging. The study interviewed a large cohort of senior people (70 years of age and over) in 1997-1998 (WAVE II) and 1999-2000 (WAVE III), respectively, and measured their health conditions, living conditions, family situations, health service utilizations, among others. Here our objective is to examine the predictive relationship between health-related events in earlier years and health outcomes in later years, which can be formulated as a multivariate regression problem.

There are $n = 3988$ common subjects who participated in both WAVE II and WAVE III interviews. After data pre-processing (Luo et al., 2018), $p = 294$ health risk and behavior measurements in WAVE II are treated as predictors, and $q = 41$ health outcomes in WAVE III are treated as multivariate responses. The response variables are binary indicators, characterizing various cognitive, sensational, social, and life quality outcomes, among others. Over 20% of the response data entries are missing. The predictors are multi-view, including housing condition ($\mathbf{X}_1$ with $p_1 = 38$), family structure/status ($\mathbf{X}_2$ with $p_2 = 60$), daily activity ($\mathbf{X}_3$ with $p_3 = 40$), prior medical condition ($\mathbf{X}_4$ with $p_4 = 114$), and medical procedure since last interview ($\mathbf{X}_5$ with $p_5 = 40$). We thus apply the proposed iRRR method to perform the regression analysis. As a comparison, we also implement gRRR (Luo et al., 2018), and both classical and sparse logistic regression methods using the R package `glmnet`, denoted as glm and glmnet, respectively.

We use a random-splitting procedure to evaluate the performance of different methods. More specifically, each time we randomly select $n_{tr} = 3000$ subjects as training samples and the remaining $n_{te} = 988$ subjects as testing samples. For each method, we use 5-fold CV on the training samples to select tuning parameters, and apply the method to all the training data with the selected tuning parameters to yield its coefficient estimate. The performance of each method is measured by the average deviance between the observed true response values and the estimated probabilities, defined as

$$\text{Average Deviance} = \frac{-2 \sum_{i=1}^{n_{te}} \sum_{j=1}^{q} \{y_{ij} \log \widehat{p}_{ij} + (1 - y_{ij}) \log(1 - \widehat{p}_{ij})\} \delta_{ij}}{\sum_{i=1}^{n_{te}} \sum_{j=1}^{q} \delta_{ij}},$$

where δ_{ij} is an indicator of whether y_{ij} is observed. We also calculate the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve for each outcome variable. This procedure is repeated 100 times and the results are averaged.

In terms of the average deviance, iRRR and glmnet yield very similar results (with mean 0.77 and standard deviation 0.01), and both substantially outperform gRRR (with mean 0.83 and standard deviation 0.01) and glm (fails due to a few singular outcomes). The out-sample AUCs for different response variables are shown in Figure 4. The response variables are sorted based on their missing rates from large (over 70%) to small (about 13%). Again, the performance of iRRR is comparable to that of glmnet. The iRRR tends to have a slight advantage over glmnet for responses with high missing rates. This could be due to the fact that iRRR can borrow information from other responses while the univariate glmnet cannot.

To understand the impact of different views on prediction, we produce heatmaps of the estimated coefficient matrices in Figure 5 (glm is omitted due to its poor performance). The estimates from iRRR and glmnet show quite similar patterns: it appears that the family structure/status group and the daily activity group have the most predictive power, and the variables within these two groups contribute to the prediction in a collective way. As for the other three views, iRRR yields heavily shrunk coefficient estimates, while glmnet yields very sparse estimates. These agreements partly explain the similarity of the two methods in their prediction performance. In contrast, the gRRR method tries to learn a globally low-rank structure rather than a view-specific structure; consequently, it yields a less parsimonious solution with less competitive prediction performance. Therefore, our results
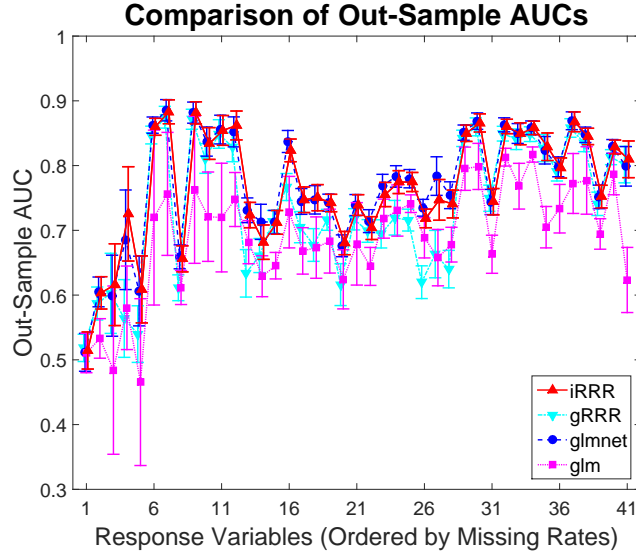
20

Figure 4: LSOA data analysis. The mean and standard deviation (error bar) of AUC for each response variable over 100 random-splitting procedures. The responses, from left to right, are ordered by missing rates from large to small.

indicate that generally knowing the family structure/status and daily activity measurements, the information on housing condition, prior medical conditions, and medical procedures do not provide much new contribution to the prediction of health outcomes on cognition, sensation, social behavior, life quality, among others.

7 Discussion

With multi-view predictor/feature sets, it is likely that some of the views are irrelevant to the prediction of the outcomes, and the features within a relevant view may be highly correlated and hence contribute to the prediction collectively rather than sparsely. When dealing with such problem, the two commonly used methodologies, i.e., sparse methods and low-rank methods, both have shortcomings. The joint extraction of latent features from each view in a supervised fashion offers a better solution; indeed, this is what iRRR strives to achieve.

There are many directions for future research. For conducting simultaneous view selection and within-view subspace selection, the proposed cNNP scheme can be extended to a general
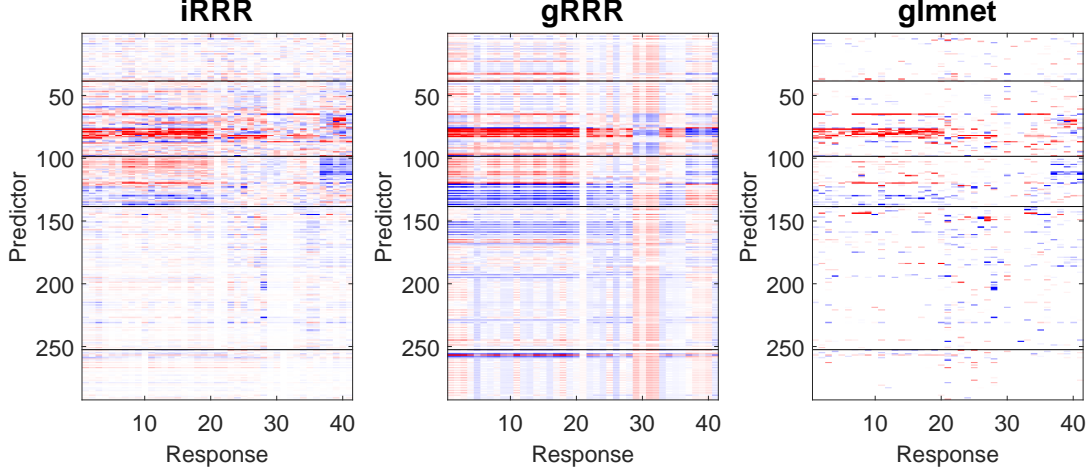
Figure 5: LSOA data analysis. The heat maps of the coefficient matrices estimated from different methods. The predictors fall into 5 groups, namely, housing condition, family status, daily activity, prior medical condition, and change in medical procedure since last interview, from top to bottom separated by horizontal black lines. For visualization purpose, we also sort the responses based on their grouping structure (e.g., cognition, sensation, social behavior, and life quality).

composite singular value penalization scheme, $\lambda \sum_{k=1}^{K} w_k \rho_{\mathcal{O}} \left(\sum_{j=1}^{p_k \wedge q} \rho_{\mathcal{I}} \left(\sigma(\mathbf{B}_k, j) \right) \right)$, where $\rho_{\mathcal{I}}$ is an inner penalty function for inducing sparsity among the singular values of each $\mathbf{B}_k$, and $\rho_{\mathcal{O}}$ is an outer penalty function for enforcing sparsity among the $\mathbf{B}_k$ matrices. For example, the family of bridge penalties (Huang et al., 2008) can be used in both inner and outer penalization. Incorporating sparse within-view variable selection to iRRR could also be fruitful; one way to achieve this is to use an additive penalty form of cNNP and grLasso. Moreover, it is possible to combine iRRR with a covariate-adjusted (inverse) covariance estimation method (Rothman et al., 2010), to jointly estimate the mean and covariance structures. Another pressing problem is to generalize iRRR to handle heterogeneous data, as in practice data may be count-valued, interval-valued, or mixed of several types with substantial missing values (Luo et al., 2018). Computationally, the ADMM algorithm can be coupled with a Majorization-Minimization algorithm to handle these cases.

Acknowledgement

A   Details on Theoretical Analysis

A.1   On the Restricted Eigenvalue Condition

To specify the restricted set $\mathcal{C}$, we need some additional constructions. For each $\mathbf{B}_{0k} \in \mathbb{R}^{p_k \times q}$ ($k = 1, \ldots, K$), let $\mathbf{B}_{0k} = \mathbf{U}_k \mathbf{D}_k \mathbf{V}_k^{\mathrm{T}}$ be its full SVD, where $\mathbf{U}_k \in \mathbb{R}^{p_k \times p_k}$, $\mathbf{V}_k \in \mathbb{R}^{q \times q}$ satisfy $\mathbf{U}_k^{\mathrm{T}} \mathbf{U}_k = \mathbf{I}_{p_k}$ and $\mathbf{V}_k^{\mathrm{T}} \mathbf{V}_k = \mathbf{I}_q$. For each $r \in \{1, 2, \cdots, m_k\}$, where $m_k = p_k \wedge q$, let $\mathbf{U}_k^r$, $\mathbf{V}_k^r$ be the submatrices of singular vectors associated with the top r singular values of $\mathbf{B}_{0k}$. Define the following subspaces of $\mathbb{R}^{p_k \times q}$:

$$
\begin{aligned}
\mathcal{A}(\mathbf{U}_k^r, \mathbf{V}_k^r) &= \{\Delta_k \in \mathbb{R}^{p_k \times q}; \mathrm{row}(\Delta_k) \subset \mathbf{V}_k^r, \mathrm{col}(\Delta_k) \subset \mathbf{U}_k^r\}, \\
\mathcal{B}(\mathbf{U}_k^r, \mathbf{V}_k^r) &= \{\Delta_k \in \mathcal{R}^{p_k \times q}; \mathrm{row}(\Delta_k) \perp \mathbf{V}_k^r, \mathrm{col}(\Delta_k) \perp \mathbf{U}_k^r\},
\end{aligned}
$$

where $\mathrm{row}(\Delta_k)$ and $\mathrm{col}(\Delta_k)$ denote the row space and column space of Δ_k, respectively. We may adopt the shorthand notation $\mathcal{A}_k^r$ and $\mathcal{B}_k^r$ when no confusion arises. Let $\mathcal{P}_{\mathcal{B}_k^{r_k}}$ denote the projection operator onto the subspace $\mathcal{B}_k^{r_k}$, and define $\Delta_k'' = \mathcal{P}_{\mathcal{B}_k^{r_k}}(\Delta_k)$ and $\Delta_k' = \Delta_k - \Delta_k''$. We now define the restricted set

$$
\begin{aligned}
&\mathcal{C}(r_1, \ldots, r_K; \delta) \\
&= \left\{ \Delta \in \mathbb{R}^{p \times q}; \|\Delta\|_{\mathbb{F}} \geq \delta, \sum_{k=1}^{K} w_k \|\Delta_k''\|_\star \leq \sum_{k=1}^{K} \left\{ 3 w_k \|\Delta_k'\|_\star + 4 w_k \sum_{j=r_k+1}^{m_k} \sigma(\mathbf{B}_{0k}, j) \right\} \right\}. \quad (13)
\end{aligned}
$$

where δ is a tolerance parameter and $w_k = \sigma(\mathbf{X}_k, 1)\{\sqrt{q} + \sqrt{r(\mathbf{X}_k)}\}/n$, as defined in (4). We refer to Negahban and Wainwright (2011) and Negahban et al. (2012) for examples of the restricted set, including the cases of Lasso, grLasso and NNP.

A.2 Special Cases of Theorem 4.2

The results on iRRR in Theorem 4.2 can specialize into oracle inequalities of several existing regularized estimation methods, such as NNP, MTL and Lasso. We discuss some examples below; to focus on the main message, we only focus on the settings of exact low rank or exact sparsity. First consider the NNP method defined in (2), which corresponds to the special case of $K = 1$ and $w_k = 1$ in iRRR. The restricted set in (13) becomes

$$\mathcal{C}(r_0) = \{\Delta \in \mathbb{R}^{p \times q}; \|\Delta''\|_\star \leq 3\|\Delta'\|_\star\},$$

where $\Delta'' = \mathcal{P}_{\mathcal{B}_0^{r_0}}(\Delta)$ and $\Delta' = \Delta - \Delta''$. Theorem 4.2 then implies that under the RE condition with $\kappa(\mathbf{X}) > 0$ over $\mathcal{C}(r_0)$, if we choose $\lambda = 2\tau(1 + \theta)\sigma(\mathbf{X}, 1)\{\sqrt{q} + \sqrt{r(\mathbf{X})}\}$, then with probability at least $1 - \exp[-\theta^2\{q + r(\mathbf{X})\}/2]$, it holds that

$$\|\widehat{\mathbf{B}} - \mathbf{B}_0\|_{\mathbb{F}}^2 \preceq \frac{\tau^2}{\kappa(\mathbf{X})^2} \frac{\{\sqrt{q} + \sqrt{r(\mathbf{X})}\}^2 r_0}{n}.$$

This bound recovers the results on NNP in the literature; see, e.g., Negahban and Wainwright (2011). Next, consider the MTL setting, which corresponds to $p_k = 1$ and $p = K$ in iRRR. Write $\mathbf{B}_0 = (\mathbf{b}_{01}^{\mathrm{T}}, \ldots, \mathbf{b}_{0p}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{p \times q}$, and $\mathcal{S} = \{j; \|\mathbf{b}_{0j}\|_2 \neq 0\}$. The restricted set becomes

$$\mathcal{C}(\mathcal{S}) = \left\{\Delta = (\Delta_1^{\mathrm{T}}, \ldots, \Delta_p^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{p \times q}; \sum_{k \in \mathcal{S}^c} \|\Delta_k\|_2 \leq 3 \sum_{k \in \mathcal{S}} \|\Delta_k\|_2\right\}.$$

By choosing $\lambda \propto \tau\sqrt{\log p/q}$, Theorem 4.2 yields the high probability bound

$$\|\widehat{\mathbf{B}} - \mathbf{B}_0\|_{\mathbb{F}}^2 \preceq \frac{\tau^2}{\kappa(\mathbf{X})^2} \frac{(\log p + q) \cdot |\mathcal{S}|}{n},$$

where $|\mathcal{S}|$ is the cardinality of $\mathcal{S}$. The same bound can be obtained from results in Lounici et al. (2011) on more general setting of MTL, or from results in Negahban et al. (2012) on grLasso by vectorizing the MTL problem here into a univariate-response regression. Another example is Lasso, which corresponds to $q = 1$ and $K = p$ in iRRR. It is seen that the model becomes $\mathbf{y} = \mathbf{X}\mathbf{b}_0 + \mathbf{e}$, and the cNNP degenerates to the ℓ_1-norm of a coefficient vector $\mathbf{b} \in \mathbb{R}^p$. Let $\mathcal{S} = \{j; b_{0j} \neq 0\}$, then the restricted set becomes

$$\mathcal{C}(\mathcal{S}) = \left\{\Delta = (\Delta_1, \ldots, \Delta_p)^{\mathrm{T}} \in \mathbb{R}^p; \sum_{k \in \mathcal{S}^c} |\Delta_k| \leq 3 \sum_{k \in \mathcal{S}} |\Delta_k|\right\}.$$

Theorem 4.2 implies that by choosing $\lambda \propto \tau\sqrt{c \log p}$,

$$\|\widehat{\mathbf{b}} - \mathbf{b}_0\|_2^2 \preceq \frac{\tau^2}{\kappa(\mathbf{X})^2} \frac{\log p \cdot |\mathcal{S}|}{n}$$

holds with probability at least $1 - p^{1-c}$, which is a well-known result in the literature.

A.3 Proof of Theorem 4.1 and a Corollary on the Estimation Error

Proof of Theorem 4.1. By definition,

$$\|\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}}\|_{\mathbb{F}}^2 + 2\lambda \sum_{k=1}^{K} \sigma(\mathbf{X}_k, 1)(\sqrt{q} + \sqrt{r(\mathbf{X}_k)})\|\mathbf{B}_k\|_\star$$
$$\leq \|\mathbf{Y} - \mathbf{X}\mathbf{C}\|_{\mathbb{F}}^2 + 2\lambda \sum_{k=1}^{K} \sigma(\mathbf{X}_k, 1)(\sqrt{q} + \sqrt{r(\mathbf{X}_k)})\|\mathbf{C}_k\|_\star,$$

which leads to

$$\|\mathbf{X}\widehat{\mathbf{B}} - \mathbf{X}\mathbf{B}_0\|_{\mathbb{F}}^2 \leq \|\mathbf{X}\mathbf{C} - \mathbf{X}\mathbf{B}_0\|_{\mathbb{F}}^2 + 2\lambda \sum_{k=1}^{K} \sigma(\mathbf{X}_k, 1)(\sqrt{q} + \sqrt{r(\mathbf{X}_k)})\|\mathbf{C}_k\|_\star$$
$$+ 2\langle \mathbf{X}^{\mathrm{T}}\mathbf{E}, \widehat{\mathbf{B}} - \mathbf{C}\rangle_{\mathbb{F}} - 2\lambda \sum_{k=1}^{K} \sigma(\mathbf{X}_k, 1)(\sqrt{q} + \sqrt{r(\mathbf{X}_k)})\|\widehat{\mathbf{B}}_k\|_\star.$$

Define an event $\mathcal{A}_k = \{\sigma(\mathbf{X}_k^{\mathrm{T}}\mathbf{E}, 1) \leq \lambda\sigma(\mathbf{X}_k, 1)(\sqrt{q} + \sqrt{r(\mathbf{X}_k)})\}$, for $k = 1, \ldots, K$. First, consider the inner product term. On the event $\cap_{k=1}^{K}\mathcal{A}_k$, we have

$$\langle \mathbf{X}^{\mathrm{T}}\mathbf{E}, \widehat{\mathbf{B}} - \mathbf{C}\rangle_{\mathbb{F}} = \mathrm{tr}\{\mathbf{E}^{\mathrm{T}}\mathbf{X}(\widehat{\mathbf{B}} - \mathbf{C})\}$$
$$= \sum_{k=1}^{K} \langle \mathbf{X}_k^{\mathrm{T}}\mathbf{E}, \widehat{\mathbf{B}}_k - \mathbf{C}_k\rangle_{\mathbb{F}}$$
$$\leq \sum_{k=1}^{K} \sigma(\mathbf{X}_k^{\mathrm{T}}\mathbf{E}, 1)\|\widehat{\mathbf{B}}_k - \mathbf{C}_k\|_\star$$
$$\leq \lambda \sum_{k=1}^{K} \sigma(\mathbf{X}_k, 1)(\sqrt{q} + \sqrt{r(\mathbf{X}_k)})\|\widehat{\mathbf{B}}_k - \mathbf{C}_k\|_\star.$$

It follows that on the event $\cap_{k=1}^{K}\mathcal{A}_k$,

$$
\begin{aligned}
\|\mathbf{X}\widehat{\mathbf{B}} - \mathbf{X}\mathbf{B}_0\|_{\mathbb{F}}^2 \leq &\|\mathbf{X}\mathbf{C} - \mathbf{X}\mathbf{B}_0\|_{\mathbb{F}}^2 + 2\lambda\sum_{k=1}^{K}\sigma(\mathbf{X}_k,1)(\sqrt{q} + \sqrt{r(\mathbf{X}_k)})\|\mathbf{C}_k\|_\star \\
&+ 2\lambda\sum_{k=1}^{K}\sigma(\mathbf{X}_k,1)(\sqrt{q} + \sqrt{r(\mathbf{X}_k)})\|\widehat{\mathbf{B}}_k - \mathbf{C}_k\|_\star \\
&- 2\lambda\sum_{k=1}^{K}\sigma(\mathbf{X}_k,1)(\sqrt{q} + \sqrt{r(\mathbf{X}_k)})\|\widehat{\mathbf{B}}_k\|_\star \\
\leq &\|\mathbf{X}\mathbf{C} - \mathbf{X}\mathbf{B}_0\|_{\mathbb{F}}^2 + 4\lambda\sum_{k=1}^{K}\sigma(\mathbf{X}_k,1)(\sqrt{q} + \sqrt{r(\mathbf{X}_k)})\|\mathbf{C}_k\|_\star, \qquad (14)
\end{aligned}
$$

where the last inequality is due to the triangle inequality.

Now we consider the probability of the event $\cap_{k=1}^{K}\mathcal{A}_k$. Let $\mathcal{P}$ be the projection matrix onto the column space of $\mathbf{X}$, and $\mathcal{P}_k$ be the projection matrix onto the column space of $\mathbf{X}_k$, for $k = 1,\ldots,K$. Because $\sigma(\mathbf{X}_k^{\mathrm{T}}\mathbf{E},1) \leq \sigma(\mathbf{X}_k,1)\sigma(\mathcal{P}_k\mathbf{E},1)$, we have

$$
\begin{aligned}
\cap_{k=1}^{K}\mathcal{A}_k =&\{\sigma(\mathbf{X}_k^{\mathrm{T}}\mathbf{E},1) \leq \lambda\sigma(\mathbf{X}_k,1)(\sqrt{q} + \sqrt{r(\mathbf{X}_k)}); k = 1,\ldots,K\} \\
\supseteq&\{\sigma(\mathcal{P}_k\mathbf{E},1) \leq \lambda(\sqrt{q} + \sqrt{r(\mathbf{X}_k)}); k = 1,\ldots,K\} \\
\equiv&\cap_{k=1}^{K}\tilde{\mathcal{A}}_k.
\end{aligned}
$$

By Lemma 3 in Bunea et al. (2011),

$$
\mathbb{P}\{(\sigma(\mathcal{P}_k\mathbf{E},1) \geq \mathbb{E}[\sigma(\mathcal{P}_k\mathbf{E},1) + \tau t]\} \leq \exp(-t^2/2),
$$

and $\mathbb{E}[\sigma(\mathcal{P}_k\mathbf{E},1)] \leq \tau(\sqrt{q} + \sqrt{r(\mathbf{X}_k)})$, for any $k = 1,\ldots,K$. Therefore,

$$
\mathbb{P}\{\cup_{k=1}^{K}\tilde{\mathcal{A}}_k^c\} \leq \sum_{k=1}^{K}\exp\{-\frac{1}{2}\theta^2(q + r(\mathbf{X}_k))\}.
$$

It then follows that

$$
\mathbb{P}\{\cap_{k=1}^{K}\mathcal{A}_k\} \geq \mathbb{P}\{\cap_{k=1}^{K}\tilde{\mathcal{A}}_k\} = 1 - \mathbb{P}\{\cup_{k=1}^{K}\tilde{\mathcal{A}}_k^c\} \geq 1 - \sum_{k=1}^{K}\exp\{-\frac{1}{2}\theta^2(q + r(\mathbf{X}_k))\}.
$$

This, together with (14), completes the proof. $\qquad\qquad\square$

Corollary. *Assume that* $\mathbf{E}$ *has i.i.d.* $N(0,\tau^2)$ *entries, and assume* $\sigma(\mathbf{X},p) > 0$. *Let* $\lambda = (1+\theta)\tau$, *with* $\theta > 0$ *arbitrary. Then with probability at least* $1 - \sum_{k=1}^{K}\exp[-\theta^2\{q + r(\mathbf{X}_k)\}/2]$,

$$
\|\widehat{\mathbf{B}} - \mathbf{B}_0\|_{\mathbb{F}}^2 \preceq \tau^2(1+\theta)^2\sum_{k=1}^{K}\frac{\Lambda(\mathbf{Z}_k,1)}{\Lambda(\mathbf{Z},p)^2}\frac{\{\sqrt{q} + \sqrt{r(\mathbf{X}_k)}\}^2 r_{0k}}{n},
$$

where "$\preceq$*" means the inequality holds up to some multiplicative constant.*

The corollary shows that the estimation error rate for iRRR is $\tau^2 \sum_{k=1}^{K} \{q + r(\mathbf{X}_k)\} r_{0k}/n$. This is potentially better than $\tau^2 \{q + r(\mathbf{X})\} r_0/n$, the rate achieved by the NNP estimator under the same conditions (Bunea et al., 2011); for example, when $r(\mathbf{X}) = \sum_k r(\mathbf{X}_k)$ and $r(\mathbf{B}_0) = \sum_k r_{0k}$.

Proof of Corollary. From the proof of Theorem 4.1,

$$
\begin{aligned}
\|\mathbf{X}\widehat{\mathbf{B}} - \mathbf{X}\mathbf{B}_0\|_{\mathbb{F}}^2 \leq & \|\mathbf{X}\mathbf{C} - \mathbf{X}\mathbf{B}_0\|_{\mathbb{F}}^2 \\
& + 2\lambda \sum_{k=1}^{K} \sigma(\mathbf{X}_k, 1)(\sqrt{q} + \sqrt{r(\mathbf{X}_k)})\{\|\mathbf{C}_k\|_{\star} + \|\widehat{\mathbf{B}}_k - \mathbf{C}_k\|_{\star} - \|\widehat{\mathbf{B}}_k\|_{\star}\}.
\end{aligned}
$$

With the results in the proof of their Theorem 12 in Bunea et al. (2011), we have

$$
\begin{aligned}
& \|\mathbf{X}\widehat{\mathbf{B}} - \mathbf{X}\mathbf{B}_0\|_{\mathbb{F}}^2 - \|\mathbf{X}\mathbf{C} - \mathbf{X}\mathbf{B}_0\|_{\mathbb{F}}^2 \\
\leq & 4\lambda \sum_{k=1}^{K} \sigma(\mathbf{X}_k, 1)(\sqrt{q} + \sqrt{r(\mathbf{X}_k)})\sqrt{3r(\mathbf{C}_k)}\|\widehat{\mathbf{B}}_k - \mathbf{C}_k\|_{\mathbb{F}} \\
\leq & 4\lambda \sqrt{3 \sum_{k=1}^{K} \sigma(\mathbf{X}_k, 1)^2(\sqrt{q} + \sqrt{r(\mathbf{X}_k)})^2 r(\mathbf{C}_k)} \sqrt{\sum_{k=1}^{K} \|\widehat{\mathbf{B}}_k - \mathbf{C}_k\|_{\mathbb{F}}^2} \\
\leq & \frac{4\sqrt{3}\lambda}{\sigma(\mathbf{X}, p)} \sqrt{\sum_{k=1}^{K} \sigma(\mathbf{X}_k, 1)^2(\sqrt{q} + \sqrt{r(\mathbf{X}_k)})^2 r(\mathbf{C}_k)} \|\mathbf{X}\widehat{\mathbf{B}} - \mathbf{X}\mathbf{C}\|_{\mathbb{F}} \\
\leq & \frac{1}{2}\|\mathbf{X}\widehat{\mathbf{B}} - \mathbf{X}\mathbf{C}\|_{\mathbb{F}}^2 + \frac{24\lambda^2}{\sigma(\mathbf{X}, p)^2} \left(\sum_{k=1}^{K} \sigma(\mathbf{X}_k, 1)^2(\sqrt{q} + \sqrt{r(\mathbf{X}_k)})^2 r(\mathbf{C}_k)\right).
\end{aligned}
$$

It follows that

$$
\|\mathbf{X}\widehat{\mathbf{B}} - \mathbf{X}\mathbf{B}_0\|_{\mathbb{F}}^2 \leq 3\|\mathbf{X}\mathbf{C} - \mathbf{X}\mathbf{B}_0\|_{\mathbb{F}}^2 + \frac{48\lambda^2}{\sigma(\mathbf{X}, p)^2} \left(\sum_{k=1}^{K} \sigma(\mathbf{X}_k, 1)^2(\sqrt{q} + \sqrt{r(\mathbf{X}_k)})^2 r(\mathbf{C}_k)\right).
$$

Taking $\mathbf{C} = \mathbf{B}_0$ leads to the bound for the prediction error

$$
\|\mathbf{X}\widehat{\mathbf{B}} - \mathbf{X}\mathbf{B}_0\|_{\mathbb{F}}^2 \preceq \tau^2 \sum_{k=1}^{K} \frac{\Lambda(\mathbf{Z}_k, 1)}{\Lambda(\mathbf{Z}, p)}(\sqrt{q} + \sqrt{r(\mathbf{X}_k)})^2 r_{0k}.
$$

Then by using the fact that $\sigma(\mathbf{X}, p) > 0$ we get the claimed bound.

$\square$

A.4   Proof of Theorem 4.2

Proof of Theorem 4.2. By definition, we have

$$\frac{1}{2n}\|\mathbf{Y}-\mathbf{X}\widehat{\mathbf{B}}\|_{\mathbb{F}}^2 + \lambda\sum_{k=1}^{K} w_k\|\widehat{\mathbf{B}}_k\|_{\star} \le \frac{1}{2n}\|\mathbf{Y}-\mathbf{X}\mathbf{B}_0\|_{\mathbb{F}}^2 + \lambda\sum_{k=1}^{K} w_k\|\mathbf{B}_{0k}\|_{\star},$$

which leads to

$$\frac{1}{2n}\|\mathbf{X}\Delta\|_{\mathbb{F}}^2 \le \lambda\sum_{k=1}^{K} w_k(\|\mathbf{B}_{0k}\|_{\star} - \|\widehat{\mathbf{B}}_k\|_{\star}) + \frac{1}{n}\langle\mathbf{E},\mathbf{X}\Delta\rangle_{\mathbb{F}}, \tag{15}$$

where $\Delta = \widehat{\mathbf{B}} - \mathbf{B}_0$.

Firstly, we verify that Δ belongs to the restricted set defined in (13) so that the RE condition can be applied. Consider the first term on the right hand side of (15). With the projection operators defined in Appendix A.1, we have that

$$\|\mathcal{P}_{\mathcal{A}_k^{r_k}}(\mathbf{B}_{0k}) + \Delta_k''\|_{\star} = \|\mathcal{P}_{\mathcal{A}_k^{r_k}}(\mathbf{B}_{0k})\|_{\star} + \|\Delta_k''\|_{\star},$$

and

$$\begin{aligned}
\|\widehat{\mathbf{B}}_k\|_{\star} &= \|\mathcal{P}_{\mathcal{A}_k^{r_k}}(\mathbf{B}_{0k}) + \Delta_k'' + \mathcal{P}_{\mathcal{B}_k^{r_k}}(\mathbf{B}_{0k}) + \Delta_k'\| \\
&\ge \|\mathcal{P}_{\mathcal{A}_k^{r_k}}(\mathbf{B}_{0k}) + \Delta_k''\|_{\star} - \|\mathcal{P}_{\mathcal{B}_k^{r_k}}(\mathbf{B}_{0k}) + \Delta_k'\|_{\star} \\
&= \|\mathcal{P}_{\mathcal{A}_k^{r_k}}(\mathbf{B}_{0k})\|_{\star} + \|\Delta_k''\|_{\star} - \|\mathcal{P}_{\mathcal{B}_k^{r_k}}(\mathbf{B}_{0k})\|_{\star} - \|\Delta_k'\|_{\star}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|\mathbf{B}_{0k}\|_{\star} - \|\widehat{\mathbf{B}}_k\|_{\star} &\le \|\mathcal{P}_{\mathcal{A}_k^{r_k}}(\mathbf{B}_{0k})\|_{\star} + \|\mathcal{P}_{\mathcal{B}_k^{r_k}}(\mathbf{B}_{0k})\|_{\star} \\
&\quad - (\|\mathcal{P}_{\mathcal{A}_k^{r_k}}(\mathbf{B}_{0k})\|_{\star} + \|\Delta_k''\|_{\star} - \|\mathcal{P}_{\mathcal{B}_k^{r_k}}(\mathbf{B}_{0k})\|_{\star} - \|\Delta_k'\|_{\star}) \\
&= 2\|\mathcal{P}_{\mathcal{B}_k^{r_k}}(\mathbf{B}_{0k})\|_{\star} + \|\Delta_k'\|_{\star} - \|\Delta_k''\|_{\star}. \tag{16}
\end{aligned}$$

We then deal with the second term on the right hand side of (15). We have

$$\begin{aligned}
\langle\mathbf{E},\mathbf{X}\Delta\rangle_{\mathbb{F}} &= \operatorname{tr}(\mathbf{E}^{\mathrm{T}}\mathbf{X}\Delta) \\
&= \sum_{k=1}^{K}\langle\mathbf{X}_k^{\mathrm{T}}\mathbf{E},\Delta_k\rangle_{\mathbb{F}} \\
&\le \sum_{k=1}^{K}\sigma(\mathbf{X}_k^{\mathrm{T}}\mathbf{E},1)\|\Delta_k\|_{\star}. \tag{17}
\end{aligned}$$

Combining results in (16) and (17), we get

$$\frac{1}{2n}\|\mathbf{X}\Delta\|_{\mathbb{F}}^2 \le \lambda \sum_{k=1}^{K} w_k (2\|\mathcal{P}_{\mathcal{B}_k^{r_k}}(\mathbf{B}_{0k})\|_\star + \|\Delta_k'\|_\star - \|\Delta_k''\|_\star) + \frac{1}{n}\sum_{k=1}^{K}\sigma(\mathbf{X}_k^{\mathrm{T}}\mathbf{E},1)\|\Delta_k\|_\star.$$

Define an event $\mathcal{A}_k = \{\sigma(\mathbf{X}_k^{\mathrm{T}}\mathbf{E},1)/n \le \lambda w_k/(1+\eta)\}$, for $k = 1, \ldots, K$, where $\eta > 0$ is an arbitrary positive number. It follows that on the event $\cap_{k=1}^{K}\mathcal{A}_k$,

$$0 \le \frac{1}{2n}\|\mathbf{X}\Delta\|_{\mathbb{F}}^2 \le \lambda \sum_{k=1}^{K} w_k (2\|\mathcal{P}_{\mathcal{B}_k^{r_k}}(\mathbf{B}_{0k})\|_\star + \|\Delta_k'\|_\star - \|\Delta_k''\|_\star)$$

$$+ \frac{\lambda}{1+\eta}\sum_{k=1}^{K} w_k \|\Delta_k\|_\star$$

$$\le \lambda \sum_{k=1}^{K} w_k \left(2\|\mathcal{P}_{\mathcal{B}_k^{r_k}}(\mathbf{B}_{0k})\|_\star + \frac{2+\eta}{1+\eta}\|\Delta_k'\|_\star - \frac{\eta}{1+\eta}\|\Delta_k''\|_\star\right).$$

Therefore, it holds that

$$\sum_{k=1}^{K} w_k \|\Delta_k''\|_\star \le \frac{2+2\eta}{\eta}\sum_{k=1}^{K} w_k \sum_{j=r_k+1}^{m_k}\sigma(\mathbf{B}_{0k},j) + \frac{2+\eta}{\eta}\sum_{k=1}^{K} w_k\|\Delta_k'\|_\star. \tag{18}$$

Taking $\eta = 1$ and assuming $\|\Delta\|_{\mathbb{F}} \ge \delta$, we see that $\Delta \in \mathcal{C}(r_1, \cdots, r_k, \delta)$. Therefore, based on the RE condition,

$$\kappa(\mathbf{X})\|\Delta\|_{\mathbb{F}}^2 \le \frac{1}{2n}\|\mathbf{X}\Delta\|_{\mathbb{F}}^2. \tag{19}$$

From (17) and on the event $\cap_{k=1}^{K}\mathcal{A}_k$, we have

$$\frac{1}{2n}\|\mathbf{X}\Delta\|_{\mathbb{F}}^2 \le \lambda \sum_{k=1}^{K} w_k (\|\mathbf{B}_{0k}\|_\star - \|\widehat{\mathbf{B}}_k\|_\star) + \frac{1}{n}\langle \mathbf{E}, \mathbf{X}\Delta\rangle_{\mathbb{F}}$$

$$\le \lambda \sum_{k=1}^{K} w_k\|\Delta_k\|_\star + \frac{\lambda}{1+\eta}\sum_{k=1}^{K} w_k\|\Delta_k\|_\star$$

$$\le \frac{2+\eta}{1+\eta}\lambda \sum_{k=1}^{K} w_k\|\Delta_k\|_\star. \tag{20}$$

From (18), we have

$$\sum_{k=1}^{K} w_k\|\Delta_k\|_\star \le \sum_{k=1}^{K} w_k\|\Delta_k'\|_\star + \sum_{k=1}^{K} w_k\|\Delta_k''\|_\star$$

$$\le \frac{2+2\eta}{\eta}\left(\sum_{k=1}^{K} w_k \sum_{j=r_k+1}^{m_k}\sigma(\mathbf{B}_{0k},j) + \sum_{k=1}^{K} w_k\|\Delta_k'\|_\star\right)$$

$$\le \frac{2+2\eta}{\eta}\left(\sum_{k=1}^{K} w_k \sum_{j=r_k+1}^{m_k}\sigma(\mathbf{B}_{0k},j) + \sum_{k=1}^{K} \sqrt{2r_k}w_k\|\Delta_k'\|_{\mathbb{F}}\right). \tag{21}$$

29

The last inequality is due to the fact that $\|\Delta\|_{\mathbb{F}} = \|\Delta'\|_{\mathbb{F}} + \|\Delta''\|_{\mathbb{F}}$.

Now, combining (19), (20) and (21), we know that on the event $\cap_{k=1}^{K}\mathcal{A}_k$, either $\|\Delta\|_{\mathbb{F}} \leq \delta$, or

$$
\kappa(\mathbf{X})\|\Delta\|_{\mathbb{F}}^2 \leq \frac{2(2+\eta)}{\eta}\lambda \left(\sum_{k=1}^{K} w_k \sum_{j=r_k+1}^{m_k} \sigma(\mathbf{B}_{0k}, j) + \sum_{k=1}^{K} \sqrt{2r_k}w_k\|\Delta_k\|_{\mathbb{F}} \right)
$$
$$
\leq \frac{2(2+\eta)}{\eta}\lambda \left(\sum_{k=1}^{K} w_k \sum_{j=r_k+1}^{m_k} \sigma(\mathbf{B}_{0k}, j) + \|\Delta\|_{\mathbb{F}}\sqrt{2\sum_{k=1}^{K} r_k w_k^2} \right).
$$

That is,

$$
\|\Delta\|_{\mathbb{F}}^2 \preceq \max \left\{ \delta^2, \frac{\lambda^2 \sum_{k=1}^{K} r_k w_k^2}{\kappa^2(\mathbf{X})}, \frac{\lambda \sum_{k=1}^{K} w_k \sum_{j=r_k+1}^{m_k} \sigma(\mathbf{B}_{0k}, j)}{\kappa(\mathbf{X})} \right\}.
$$

Lastly, from the proof of Theorem 4.1, choosing $\lambda = 2(1+\theta)\tau$ ensures that

$$
\mathbb{P}\{\cap_{k=1}^{K}\mathcal{A}_k\} \geq 1 - \sum_{k=1}^{K} \exp\{-\frac{1}{2}\theta^2(q + r(\mathbf{X}_k))\}.
$$

This completes the proof.

$\square$

B   Additional Simulation with Correlated Errors

We conduct additional simulation studies where the errors in $\mathbf{E}$ are correlated. In particular, we consider an AR(1) covariance structure with common variance 1 and autocorrelation 0.5 for the random errors in $\mathbf{E}$ in Settings 1–5. The same methods are used and the results are shown in Table 2 and Figure 6. The results are very similar to those with i.i.d. errors. A closer look reveals that the proposed iRRR method is very robust against the violation of the independent error assumption, while other methods (especially MTL and grLasso) are more sensitive.

Table 2: Simulation results for **Settings 1–4** with correlated errors. The mean and standard deviation (in parenthesis) of MSPE over 100 simulation runs are presented. In each setting, the best results are highlighted in boldface.

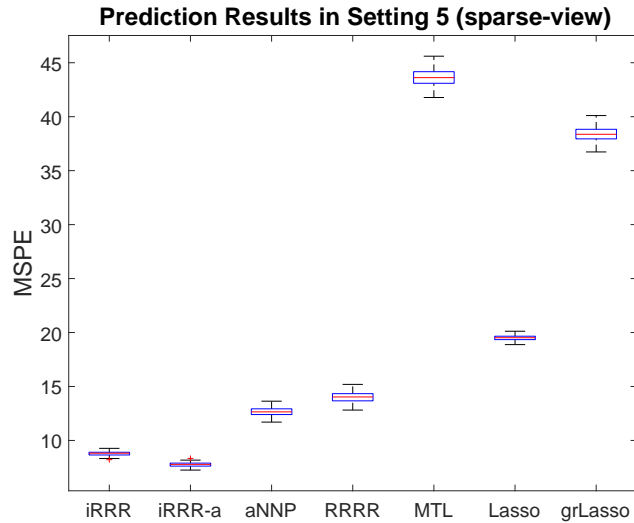|	iRRR	aNNP	RRRR	OLS
Setting 1	**7.74** (0.22)	9.11 (0.32)	10.27 (0.43)	25.14 (0.58)
Setting 2	**4.62** (0.10)	5.63 (0.18)	5.35 (0.14)	25.14 (0.58)
$(r_0 = 20)$	10.73 (0.26)	**9.10** (0.33)	10.06 (0.45)	25.17 (0.60)
Setting 3 $(r_0 = 40)$	**13.10** (0.24)	14.09 (0.33)	15.08 (0.17)	25.11 (0.60)
$(r_0 = 60)$	**14.40** (0.23)	16.43 (0.38)	15.70 (0.16)	25.16 (0.52)
$(K = 3)$	**11.03** (0.26)	17.11 (0.48)	17.63 (0.24)	43.87 (0.84)
Setting 4 $(K = 4)$	**14.12** (0.25)	25.33 (0.64)	20.97 (0.22)	68.06 (1.29)
$(K = 5)$	**16.09** (0.29)	30.78 (0.40)	23.01 (0.22)	101.81 (1.45)

Figure 6: Simulation results for **Setting 5 (sparse-view)** with correlated errors.

References

Anderson, T. W. (1951). Estimating linear restrictions on regression coefficients for multi-variate normal distributions. *Annals of Mathematical Statistics 22*, 327–351.

Boyd, S., N. Parikh, E. Chu, B. Peleato, and J. Eckstein (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning 3*(1), 1–122.

Breheny, P. and J. Huang (2009). Penalized methods for bi-level variable selection. *Statistics and Its Interface 2*(3), 369–380.

Bunea, F., Y. She, and M. Wegkamp (2012). Joint variable and rank selection for parsimonious estimation of high dimensional matrices. *The Annals of Statistics 40*(5), 2359–2388.

Bunea, F., Y. She, and M. H. Wegkamp (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics 39*(2), 1282–1309.

Cai, J.-F., E. J. Candès, and Z. Shen (2010). A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization 20*(4), 1956–1982.

Caruana, R. (1997, Jul). Multitask learning. *Machine Learning 28*(1), 41–75.

Chen, K., K.-S. Chan, and N. C. Stenseth (2012). Reduced rank stochastic regression with a sparse singular value decomposition. *Journal of the Royal Statistical Society: Series B 74*(2), 203–221.

Chen, K., H. Dong, and K.-S. Chan (2013). Reduced rank regression via adaptive nuclear norm penalization. *Biometrika 100*(4), 901–920.

Chen, K., E. A. Hoffman, I. Seetharaman, C.-L. Lin, and K.-S. Chan (2016). Linking lung airway structure to pulmonary function via composite bridge regression. *The Annals of Applied Statistics 10*(4), 1880–1906.

Chen, L. and J. Z. Huang (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association 107*(500), 1533–1545.

Cook, R. D., L. Forzani, and X. Zhang (2015). Envelopes and reduced-rank regression. *Biometrika 102*(2), 439–456.

He, B., L.-Z. Liao, D. Han, and H. Yang (2002). A new inexact alternating directions method for monotone variational inequalities. *Mathematical Programming 92*(1), 103–118.

He, B., H. Yang, and S. Wang (2000). Alternating direction method with self-adaptive penalty parameters for monotone variational inequalities. *Journal of Optimization Theory and Applications 106*(2), 337–356.

Huang, J., P. Breheny, and S. Ma (2012). A selective review of group selection in high dimensional models. *Statistical Science 27*(4), 481–499.

Huang, J., S. Ma, and C.-H. Zhang (2008). Adaptive lasso for high-dimensional regression models. *Statistica Sinica 18*, 1603–1618.

Koltchinskii, V., K. Lounici, and A. B. Tsybakov (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics 39*(5), 2302–2329.

Lee, S. and J. Z. Huang (2013). A coordinate descent mm algorithm for fast computation of sparse logistic pca. *Computational Statistics & Data Analysis 62*, 26–38.

Lee, S., J. Z. Huang, and J. Hu (2010). Sparse logistic principal components analysis for binary data. *The Annals of Applied Statistics 4*(3), 1579.

Lee, W. and Y. Liu (2012). Simultaneous multiple response regression and inverse covariance matrix estimation via penalized gaussian maximum likelihood. *Journal of Multivariate Analysis 111*, 241–255.

Li, Y., B. Nan, and J. Zhu (2015). Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics 71*(2), 354–363.

Liu, J., S. Ma, and J. Huang (2014, November). Integrative Analysis of Cancer Diagnosis Studies with Composite Penalization. *Scandinavian Journal of Statistics 41*(1), 87–103.

Lounici, K., M. Pontil, S. van de Geer, and A. B. Tsybakov (2011). Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics 39*(4), 2164–2204.

Luo, C., J. Liang, G. Li, F. Wang, C. Zhang, D. K. Dey, and K. Chen (2018). Leveraging mixed and incomplete outcomes via reduced-rank modeling. *Journal of Multivariate Analysis 167*, 378–394.

Ma, S., J. Huang, F. Wei, Y. Xie, and K. Fang (2011). Integrative analysis of multiple cancer prognosis studies with gene expression measurements. *Statistics in Medicine 30*(28), 3361–3371.

Mukherjee, A. and J. Zhu (2011). Reduced rank ridge regression and its kernel extensions. *Statistical Analysis and Data Mining 4*(6), 612–622.

Negahban, S. and M. J. Wainwright (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics 39*(2), 1069–1097.

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

Peng, J., J. Zhu, A. Bergamaschi, W. Han, D.-Y. Noh, J. R. Pollack, and P. Wang (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *The Annals of Applied Statistics 4*(1), 53–77.

Reinsel, G. C. and P. Velu (1998). *Multivariate reduced-rank regression: theory and applications.* New York: Springer.

Rothman, A. J., E. Levina, and J. Zhu (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics 19*(4), 947–962.

She, Y. (2013). Reduced rank vector generalized linear models for feature extraction. *Statistics and Its Interface 6*, 413–424.

Stanziano, D. C., M. Whitehurst, P. Graham, and B. A. Roos (2010). A review of selected longitudinal studies on aging: past findings and future directions. *Journal of the American Geriatrics Society 58*(2), 292–297.

Stone, M. (1974). Cross-validation and multinomial prediction. *Biometrika 61*(3), 509–515.

Su, Z., G. Zhu, X. Chen, and Y. Yang (2016). Sparse envelope model: efficient estimation and response variable selection in multivariate linear regression. *Biometrika 103*(3), 579–593.

Sun, T. and C.-H. Zhang (2012). Calibrated elastic regularization in matrix completion. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, USA, pp. 863–871. Curran Associates Inc.

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

Velu, R. P. (1991). Reduced rank models with two sets of regressors. *Journal of the Royal Statistical Society: Series C 40*(1), 159–170.

Yuan, M., A. Ekici, Z. Lu, and R. Monteiro (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B 69*(3), 329–346.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B 68*(1), 49–67.

Zhou, H. and L. Li (2014). Regularied matrix regression. *Journal of the Royal Statistical Society: Series B 76*(2), 463–483.

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

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