Biometrics 0, 1–16 DOI: 000
July 2018

Web-based Supplementary Materials for "Integrative Multi-View Regression: Bridging Group-Sparse and Low-Rank Models"

Gen Li

Department of Biostatistics, Columbia University

Xiaokang Liu

Department of Statistics, University of Connecticut, Storrs, CT

and

Kun Chen*

Department of Statistics, University of Connecticut, Storrs, CT

email: kun.chen@uconn.edu

This paper has been submitted for consideration for publication in *Biometrics*

Web Appendix A. Computational Algorithm and Further Extensions

A.1 *ADMM for iRRR*

For readers convenience, we first reproduce the proposed iRRR estimator defined in (3) of the main paper. It is given by

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

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. To adjust for the dimension and scale differences of $\mathbf{X}_k$s, we choose

$$w_k = \sigma(\mathbf{X}_k, 1)\{\sqrt{q}+\sqrt{r(\mathbf{X}_k)}\}/n, \tag{2}$$

based on a concentration inequality of the largest singular value of a Gaussian matrix.

Without loss of generality, we omit the weights w_k $(k = 1,\cdots,K)$ 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

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

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

$$\begin{aligned}\mathcal{D}(\mathbf{Y};\mathbf{A},\mathbf{B},\mathbf{\Lambda}) =& \frac{1}{2n}\|\mathbf{Y}-\sum_{k=1}^{K}\mathbf{X}_k\mathbf{B}_k\|_{\mathbb{F}}^2 + \lambda\sum_{k=1}^{K}\|\mathbf{A}_k\|_\star \\ &+ \sum_{k=1}^{K}\langle\mathbf{\Lambda}_k, \mathbf{A}_k-\mathbf{B}_k\rangle_{\mathbb{F}} + \frac{\rho}{2}\sum_{k=1}^{K}\|\mathbf{A}_k-\mathbf{B}_k\|_{\mathbb{F}}^2,\end{aligned} \tag{3}$$

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 $\mathcal{D}(\mathbf{Y};\mathbf{A},\mathbf{B},\mathbf{\Lambda})$ with respect to $\mathbf{A}$ and $\mathbf{B}$, respectively, while fixing everything else; the *dual* step updates $\mathbf{\Lambda}$.

Primal step: We minimize (3) 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 $\widetilde{\mathbf{\Lambda}}$ represent the estimates from the previous iteration. The optimization $\min_{\mathbf{B}} \mathcal{D}(\mathbf{Y};\widetilde{\mathbf{A}},\mathbf{B},\widetilde{\mathbf{\Lambda}})$ has a unique solution

$$\widehat{\mathbf{B}} = \left(\frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X} + \rho\mathbf{I}\right)^{-1}\left(\frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{Y} + \rho\widetilde{\mathbf{A}} + \widetilde{\mathbf{\Lambda}}\right). \tag{4}$$

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

$$\min_{\mathbf{A}_k} \quad \frac{\rho}{2}\|\mathbf{A}_k - \widehat{\mathbf{B}}_k + \frac{\widetilde{\mathbf{\Lambda}}_k}{\rho}\|_{\mathbb{F}}^2 + \lambda\|\mathbf{A}_k\|_\star, \tag{5}$$

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 $\widehat{\mathbf{B}}_k - \widetilde{\mathbf{\Lambda}}_k/\rho$, 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 (5) is

$$\widehat{\mathbf{A}}_k = \mathbf{U}_k\mathcal{S}(\mathbf{D}_k, \frac{\lambda}{\rho})\mathbf{V}_k^{\mathrm{T}}, \tag{6}$$

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 $\mathbf{\Lambda}_k$ are updated by

$$\widehat{\mathbf{\Lambda}}_k = \widetilde{\mathbf{\Lambda}}_k + \rho(\widehat{\mathbf{A}}_k - \widehat{\mathbf{B}}_k). \tag{7}$$

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

$$r_{primal} = \|\widehat{\mathbf{A}} - \widehat{\mathbf{B}}\|_{\mathbb{F}}, \quad r_{dual} = \rho\|\widehat{\mathbf{B}} - \widetilde{\mathbf{B}}\|_{\mathbb{F}}. \tag{8}$$

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 (4) and update $\mathbf{A}$ by (6);
- Dual step: update $\mathbf{\Lambda}$ by (7);
- Calculate the primal and dual residuals in (8);
- (Optional) Increase ρ by a small amount, e.g., $\rho \leftarrow 1.1\rho$.

end while

The tuning parameter λ in (1) 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.

A.2 *Handling Non-Gaussian and Incomplete Response*

When the responses are non-Gaussian, we substitute the squared loss function in (1) 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 $\mathbf{\Theta} = \mathbf{1}\boldsymbol{\mu}^{\mathrm{T}} + \mathbf{X}\mathbf{B}$. The minimization of $\mathcal{D}(\mathbf{Y}; \boldsymbol{\mu}, \mathbf{A}, \mathbf{B}, \mathbf{\Lambda})$ 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

$$\log L(\mathbf{Y}, \mathbf{\Theta}) = \sum_{i=1}^{n} \sum_{j=1}^{q} \log h\left((2y_{ij} - 1)\theta_{ij}\right), \tag{9}$$

where θ_{ij} is the (i, j)th entry of $\mathbf{\Theta}$ 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

$$-\log h(\eta) \leqslant -\log h(\eta_0) - 2\{1 - h(\eta_0)\}^2 + \frac{1}{8}\left[\eta - \eta_0 - 4\{1 - h(\eta_0)\}\right]^2. \tag{10}$$

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 (10) to (9), we have

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

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 $-\log L(\mathbf{Y}, \mathbf{\Theta})$ is majorized by the squared function $1/8\|\mathbf{Y}^\star - \mathbf{\Theta}\|_{\mathbb{F}}^2$, 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 $1/8\sum_{(i,j)\in\mathcal{O}}(y^{\star}_{ij}-\theta_{ij})^2$ as before, and then further majorize it as $1/8\sum_{(i,j)\in\mathcal{O}}(y^{\star}_{ij}-\theta_{ij})^2 + 1/8\sum_{(i,j)\in\mathcal{M}}(\widetilde{\theta}_{ij}-\theta_{ij})^2$. By collecting y_{ij} or $y^{\star}_{ij}$ 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.

A.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 (1) (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 (5) 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 4 of the main paper.

Web Appendix B. 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

$$\begin{aligned}&\mathcal{C}(r_1, \ldots, r_K; \delta)\\ &= \left\{ \Delta \in \mathbb{R}^{p\times q}; \|\Delta\|_{\mathbb{F}} \geqslant \delta, \sum_{k=1}^K w_k \|\Delta_k''\|_\star \leqslant \sum_{k=1}^K \{3w_k\|\Delta_k'\|_\star + 4w_k \sum_{j=r_k+1}^{m_k} \sigma(\mathbf{B}_{0k}, j)\} \right\}. \end{aligned} \tag{11}$$

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) of the main paper. We refer to Negahban and Wainwright (2011) and Negahban et al. (2012) for examples of the restricted set, including the cases of Lasso, group Lasso and nuclear norm penalized regression.

Web Appendix C. Special Cases of Theorem 2

The results on iRRR in Theorem 2 of the main paper 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) of the main paper, which corresponds to the special case of $K = 1$ and $w_k = 1$ in iRRR. The restricted set in (11) becomes

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

where $\Delta'' = \mathcal{P}_{\mathcal{B}_0^{r_0}}(\Delta)$ and $\Delta' = \Delta - \Delta''$. Theorem 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

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

By choosing $\lambda \propto \tau\sqrt{\log p/q}$, Theorem 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

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

Theorem 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.

Web Appendix D. Proofs

D.1 *Proof of Theorem 1 and a Corollary on the Estimation Error*

Proof. [Proof of Theorem 1] By definition,

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

which leads to

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

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

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

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

$$\begin{aligned}\|\mathbf{X}\widehat{\mathbf{B}} - \mathbf{X}\mathbf{B}_0\|_{\mathbb{F}}^2 \leqslant & \|\mathbf{X}\mathbf{C} - \mathbf{X}\mathbf{B}_0\|_{\mathbb{F}}^2 + 2\lambda \sum_{k=1}^{K} \sigma(\mathbf{X}_k, 1)(\sqrt{q} + \sqrt{r(\mathbf{X}_k)})\|\mathbf{C}_k\|_\star \\ & + 2\lambda \sum_{k=1}^{K} \sigma(\mathbf{X}_k, 1)(\sqrt{q} + \sqrt{r(\mathbf{X}_k)})\|\widehat{\mathbf{B}}_k - \mathbf{C}_k\|_\star \\ & - 2\lambda \sum_{k=1}^{K} \sigma(\mathbf{X}_k, 1)(\sqrt{q} + \sqrt{r(\mathbf{X}_k)})\|\widehat{\mathbf{B}}_k\|_\star \\ \leqslant & \|\mathbf{X}\mathbf{C} - \mathbf{X}\mathbf{B}_0\|_{\mathbb{F}}^2 + 4\lambda \sum_{k=1}^{K} \sigma(\mathbf{X}_k, 1)(\sqrt{q} + \sqrt{r(\mathbf{X}_k)})\|\mathbf{C}_k\|_\star, \qquad (12)\end{aligned}$$

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 $\sigma(\mathbf{X}_k^{\mathrm{T}}\mathbf{E}, 1) \leqslant \sigma(\mathbf{X}_k, 1)\sigma(\mathcal{P}_k\mathbf{E}, 1)$, we have

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

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

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

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

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

It then follows that

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

This, together with (12), completes the proof.

COROLLARY 1: *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.*

Corollary 1 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. [Proof of Corollary 1] From the proof of Theorem 1,

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

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

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

It follows that

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

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.

D.2 *Proof of Theorem 2*

Proof. [Proof of Theorem 2] By definition, we have

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

which leads to

$$\frac{1}{2n}\|\mathbf{X}\Delta\|_{\mathbb{F}}^2 \leqslant \lambda \sum_{k=1}^{K} w_k(\|\mathbf{B}_{0k}\|_\star - \|\widehat{\mathbf{B}}_k\|_\star) + \frac{1}{n}\langle \mathbf{E}, \mathbf{X}\Delta \rangle_{\mathbb{F}}, \tag{13}$$

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

Firstly, we verify that Δ belongs to the restricted set defined in (11) so that the RE condition can be applied. Consider the first term on the right hand side of (13). With the projection operators defined in Web Appendix B, 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

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

Therefore,

$$\begin{aligned}\|\mathbf{B}_{0k}\|_\star - \|\widehat{\mathbf{B}}_k\|_\star &\leqslant \|\mathcal{P}_{\mathcal{A}_k^{r_k}}(\mathbf{B}_{0k})\|_\star + \|\mathcal{P}_{\mathcal{B}_k^{r_k}}(\mathbf{B}_{0k})\|_\star \\ &\quad - (\|\mathcal{P}_{\mathcal{A}_k^{r_k}}(\mathbf{B}_{0k})\|_\star + \|\Delta_k''\|_\star - \|\mathcal{P}_{\mathcal{B}_k^{r_k}}(\mathbf{B}_{0k})\|_\star - \|\Delta_k'\|_\star) \\ &= 2\|\mathcal{P}_{\mathcal{B}_k^{r_k}}(\mathbf{B}_{0k})\|_\star + \|\Delta_k'\|_\star - \|\Delta_k''\|_\star. \qquad (14)\end{aligned}$$

We then deal with the second term on the right hand side of (13). We have

$$\begin{aligned}\langle \mathbf{E}, \mathbf{X}\Delta\rangle_{\mathbb{F}} &= \mathrm{tr}(\mathbf{E}^{\mathrm{T}}\mathbf{X}\Delta) \\ &= \sum_{k=1}^{K}\langle \mathbf{X}_k^{\mathrm{T}}\mathbf{E}, \Delta_k\rangle_{\mathbb{F}} \\ &\leqslant \sum_{k=1}^{K}\sigma(\mathbf{X}_k^{\mathrm{T}}\mathbf{E}, 1)\|\Delta_k\|_\star. \qquad (15)\end{aligned}$$

Combining results in (14) and (15), we get

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

Define an event $\mathcal{A}_k = \{\sigma(\mathbf{X}_k^{\mathrm{T}}\mathbf{E}, 1)/n \leqslant \lambda w_k/(1+\eta)\}$, for $k = 1, \ldots, K$, where $\eta > 0$ is an

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

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

Therefore, it holds that

$$\sum_{k=1}^K w_k \|\Delta_k''\|_\star \leqslant \frac{2+2\eta}{\eta}\sum_{k=1}^K w_k \sum_{j=r_k+1}^{m_k} \sigma(\mathbf{B}_{0k}, j) + \frac{2+\eta}{\eta}\sum_{k=1}^K w_k \|\Delta_k'\|_\star. \tag{16}$$

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

$$\kappa(\mathbf{X})\|\Delta\|_{\mathbb{F}}^2 \leqslant \frac{1}{2n}\|\mathbf{X}\Delta\|_{\mathbb{F}}^2. \tag{17}$$

From (15) and on the event $\cap_{k=1}^K \mathcal{A}_k$, we have

$$\begin{aligned} \frac{1}{2n}\|\mathbf{X}\Delta\|_{\mathbb{F}}^2 &\leqslant \lambda \sum_{k=1}^K w_k (\|\mathbf{B}_{0k}\|_\star - \|\widehat{\mathbf{B}}_k\|_\star) + \frac{1}{n}\langle \mathbf{E}, \mathbf{X}\Delta\rangle_{\mathbb{F}} \\ &\leqslant \lambda \sum_{k=1}^K w_k \|\Delta_k\|_\star + \frac{\lambda}{1+\eta}\sum_{k=1}^K w_k \|\Delta_k\|_\star \\ &\leqslant \frac{2+\eta}{1+\eta}\lambda \sum_{k=1}^K w_k \|\Delta_k\|_\star. \end{aligned} \tag{18}$$

From (16), we have

$$\begin{aligned} \sum_{k=1}^K w_k \|\Delta_k\|_\star &\leqslant \sum_{k=1}^K w_k \|\Delta_k'\|_\star + \sum_{k=1}^K w_k \|\Delta_k''\|_\star \\ &\leqslant \frac{2+2\eta}{\eta}\left(\sum_{k=1}^K w_k \sum_{j=r_k+1}^{m_k} \sigma(\mathbf{B}_{0k}, j) + \sum_{k=1}^K w_k \|\Delta_k'\|_\star\right) \\ &\leqslant \frac{2+2\eta}{\eta}\left(\sum_{k=1}^K w_k \sum_{j=r_k+1}^{m_k} \sigma(\mathbf{B}_{0k}, j) + \sum_{k=1}^K \sqrt{2r_k} w_k \|\Delta_k'\|_{\mathbb{F}}\right). \end{aligned} \tag{19}$$

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

Now, combining (17), (18) and (19), we know that on the event $\cap_{k=1}^K \mathcal{A}_k$, either $\|\Delta\|_{\mathbb{F}} \leqslant \delta$,

or

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

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 1, choosing $\lambda = 2(1+\theta)\tau$ ensures that

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

This completes the proof.

Web Appendix E. 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 presented in the main paper. The same methods are used and the results are shown in Web Table 1 and Web Figure 1. The results are very similar to those with i.i.d. errors in the main paper. 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.

[Web Table. 1 about here.]

[Web Figure 1 about here.]

References

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations*

and Trends® in Machine Learning **3,** 1–122.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

23 July 2018

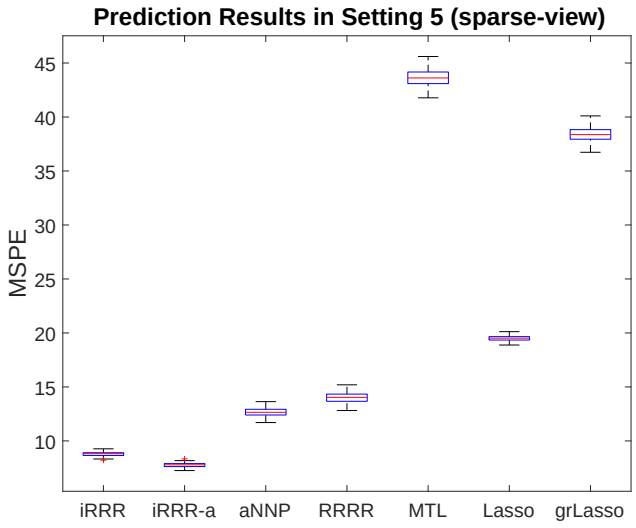

Figure 1 Simulation results for **Setting 5 (sparse-view)** with correlated errors.

Table 1 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)
Setting 3	($r_0 = 20$)	10.73 (0.26)	**9.10** (0.33)	10.06 (0.45)	25.17 (0.60)
	($r_0 = 40$)	**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)
Setting 4	($K = 3$)	**11.03** (0.26)	17.11 (0.48)	17.63 (0.24)	43.87 (0.84)
	($K = 4$)	**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)