Journal of Multivariate Analysis 167 (2018) 378–394

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Leveraging mixed and incomplete outcomes via reduced-rank modeling

Chongliang Luo[a], Jian Liang[b], Gen Li[c], Fei Wang[d], Changshui Zhang[d], Dipak K. Dey[a], Kun Chen[a,*]

[a] Department of Statistics, University of Connecticut, Storrs, CT, United States
[b] Department of Automation, Tsinghua University, Beijing, China
[c] Department of Biostatistics, Columbia University, New York, United States
[d] Department of Healthcare Policy and Research, Weill Cornell Medical School, Cornell University, New York, United States

ARTICLE INFO

Article history:
Received 29 August 2017
Available online 2 July 2018

AMS subject classifications:
62H20
62H25

Keywords:
Generalized linear model
Integrative learning
Missing data
Multivariate regression

ABSTRACT

Multivariate outcomes with multivariate features of possibly high dimension are routinely produced in various fields. In many real-world problems, the collected outcomes are of mixed types, including continuous measurements, binary indicators and counts, and a substantial proportion of values may also be missing. Regardless of their types, these mixed outcomes are often interrelated, representing diverse reflections or views of the same underlying data generation mechanism. As such, an integrative multivariate model can be beneficial. We develop a mixed-outcome reduced-rank regression, which effectively enables information sharing among different prediction tasks. Our approach integrates mixed and partially observed outcomes belonging to the exponential dispersion family, by assuming that all the outcomes are associated through a shared low-dimensional subspace spanned by the features. A general singular value regularized criterion is proposed, and we establish a non-asymptotic performance bound for the proposed estimators in the context of supervised learning with mixed outcomes from an exponential family and under a general sampling scheme of missing data. An iterative singular value thresholding algorithm is developed for optimization with convergence guarantee. The effectiveness of our approach is demonstrated by simulation studies and an application on predicting health-related outcomes in longitudinal studies of aging.

1. Introduction

Multivariate outcomes/responses, or measurements of diverse and yet interrelated characteristics pertaining to a single set of subjects, together with multivariate features/predictors of possibly high dimension, are routinely produced in various fields of scientific research as well as in our daily lives. Many associated statistical learning problems fall into the domain of multivariate regression analysis, whose main objective is to build an accurate and interpretable predictive model for the outcomes of interest. In a human lung study of asthmatics, for example, the goal is to understand how overall functions of lung are influenced by microscopic lung airway structure [12], which amounts to associating discrete clinically-determined asthma severity status or continuous asthma quality of life scores from questionnaires with high-dimensional measurements of lung airway tree from a computed tomography scan. In an adolescent health study, annual hospitalization counts due

[*] Corresponding author.
E-mail address: kun.chen@uconn.edu (K. Chen).

https://doi.org/10.1016/j.jmva.2018.04.011

to various causes such as disease, accidental injury, self-inflicted injury, etc., were collected for each school district in a state [8]; the interest was to understand how the various types of health-related risks, proxied by the hospitalization counts, were related to demographics, social-economic factors, academic performances, etc. In studies of aging on elderly subjects [45], continuous measurements of health, memory and sensation scores, dichotomous measurements of various medical conditions may be well predicted by subject demographics and records of medical history, life style and social behavior.

In the aforementioned examples, several types of outcomes, e.g., continuous, binary, and count data, may all be collected from the same cohort in the same study. We refer to such a collection of outcomes as *mixed outcomes* or *outcomes of mixed types*. In general, regardless of their types of measurements, such outcomes are expected to be related, as they commonly represent diverse reflections or different views of the underlying data generation mechanism. Therefore, an integrative learning of the mixed outcomes could be highly preferable in order to enable information sharing among different prediction tasks.

However, most existing multivariate techniques are only applicable for analyzing one type of outcomes at a time. For continuous outcomes, multivariate linear regression and its extensions have been extensively studied, e.g., ridge regression for overcoming multicollinearity [24], sparse regression for variable selection [19,37,40,52], and reduced-rank regression for dimension reduction [1,4,39]. Besides rank-constrained estimation, reduced-rank models can be realized through singular value regularization [10,27,33,36,51,54]. Recently, several authors considered sparse and reduced-rank models [5,9,11,48]. As soon as we step into the territory of non-Gaussian and/or non-linear analysis, the modeling of multivariate dependency becomes much more complicated. Vector generalized linear models were extended from their univariate counterparts based on a multivariate analogue of dispersion model family distributions, in which the correlation of the outcomes is explicitly modeled by an association matrix; see [44] for a comprehensive review on related topics. Yee and Hastie [50] proposed reduced-rank vector generalized linear models (RR-VGLM). She [42] further studied RR-VGLM and proposed an iterative algorithm with convergence guarantee. However, neither of them considered incomplete data and studied the theoretical properties of RR-VGLM. Yuan et al. [51] studied semiparametric and nonparametric low-rank models. There is also a rich literature on using sufficient dimension reduction to explore multivariate association; see [14,30,31] and the references therein.

Simultaneous statistical modeling of mixed outcomes is under-explored thus far. To the best of our knowledge, most of the existing approaches attempt to model the correlation among mixed outcomes in an explicit way; a drawback of such, however, is that it may not be applicable in high-dimensional settings. Cox and Wermuth [16] and Fitzmaurice and Laird [22] considered likelihood based methods by factorizing the joint distribution as marginal and conditional distributions. Prentice and Zhao [38] and Zhao et al. [53] used generalized estimating equations [32] to handle mixtures of continuous and discrete outcomes. Indeed, direct joint modeling of mixed outcomes is challenging due to the lack of convenient multivariate distributions, even when the number of outcomes is small. Another strategy is to induce multivariate dependency through some shared latent variable, conditional on which the outcomes are then assumed to be independent [17,41].

Our particular interest here is on generalizing and leveraging a reduced-rank matrix structure for modeling mixed multivariate outcomes with multivariate features, both of which are possibly of high dimension. From the publication of the seminal work by Anderson [1] several decades ago to the current era of big data, reduced-rank models have been very attractive, especially for modeling continuous multivariate data, in which the low-rank assumption of certain coefficient matrices conveniently captures the dependencies among the variables and systematically mitigates the curse of dimensionality. In the regression context, the low-rank assumption translates into a latent variable model, implying that all the outcomes are associated with the same small set of latent variables that are themselves linear functions of the original high-dimensional features/predictors. This elegant idea brings a genuine multivariate flavor to the model and, in an implicit way, induces and takes advantage of the dependency among the outcomes.

Giving the prevalence of big data, it is appealing to explore the use of reduced-rank structure in an integrative analysis of mixed outcomes, especially when the main goal is on dimension reduction and prediction. Similar ideas recently appeared in Udell et al. [49], in which the authors mainly focused on unsupervised learning and computation. Here, our goal is to provide a comprehensive study on a *mixed-response reduced-rank generalized linear regression model* (mRRR). The main contributions of this paper and some key features of our proposed approach are outlined as follows:

(i) Our approach integrates multivariate outcomes of mixed types belonging to an exponential dispersion family, and is able to conveniently handle incomplete data records in the multivariate statistical analysis.
(ii) We study the theoretical properties of mRRR in a general high-dimensional non-asymptotic framework. Finite-sample performance bounds are established for mRRR under a general setup of incomplete and mixed outcomes from an exponential family.
(iii) We provide a general, practical modeling framework and computational implementation for analyzing high-dimensional mixed outcomes, taking into account offset terms, fixed effects of control variables, and differential dispersion of the mixed-type outcomes. Our model and implementation can be readily extended to enable robust estimation, variable selection, etc.

The rest of the paper is organized as follows. In Section 2, we propose the mRRR framework for jointly analyzing mixed outcomes. In Section 3, we establish oracle inequalities for mRRR. A unified iterative algorithm is presented in Section 4. The performance gain by mRRR over several alternative modeling strategies is demonstrated via simulations in Section 5. In Section 6, we apply mRRR to build a joint predictive model of health conditions with data from studies of aging. Some concluding remarks are provided in Section 7.

Table 1
Some common distributions in the exponential dispersion family.

Distribution	Mean	Variance	θ	ϕ	$a(\phi)$	$b(\theta)$	$c(y;\phi)$
Bernoulli(p)	p	$p(1-p)$	$\ln\{p/(1-p)\}$	1	1	$\ln(1+e^{\theta})$	0
$\mathcal{P}(\lambda)$	λ	λ	$\ln\lambda$	1	1	e^{θ}	$-\ln y!$
$\mathcal{N}(\mu,\sigma^2)$	μ	σ^2	μ	σ^2	ϕ	$\theta^2/2$	$-(y^2\phi^{-1}+\ln 2\pi)/2$
$\mathcal{G}(\alpha,\beta)$	α/β	α/β^2	$-\beta/\alpha$	$1/\alpha$	ϕ	$-\ln(-\theta)$	$\ln\{\alpha^{\alpha}y^{\alpha-1}/\Gamma(\alpha)\}$

2. Mixed-response reduced-rank regression

Let $\mathbf{Y}=(\tilde{\mathbf{y}}_1,\ldots,\tilde{\mathbf{y}}_q)=(\mathbf{y}_1,\ldots,\mathbf{y}_n)^\top\in\mathbb{R}^{n\times q}$ be the complete response matrix consisting of n mutually independent observations from q outcome/response variables. In many applications, however, $\mathbf{Y}$ may be partially observed. Let

$$\Omega=\{(i,k);\ y_{ik}\text{ is observed, }i\in\{1,\ldots,n\},k\in\{1,\ldots,q\}\}$$

be an index set collecting all the entries corresponding to the observed outcomes. Let $\widetilde{\mathbf{Y}}=P_\Omega(\mathbf{Y})$ denote the projection of $\mathbf{Y}$ onto Ω, i.e., $\tilde{y}_{ik}=y_{ik}$ for any $(i,k)\in\Omega$ and $\tilde{y}_{ik}=0$ otherwise.

We assume that each y_{ik}, i.e., the ith observation on the kth outcome variable, for any $(i,k)\in\Omega$, follows a distribution from the exponential dispersion family [25]. Specifically, the probability density function of each y_{ik} takes the form

$$f(y_{ik};\theta_{ik},\phi_k)=\exp\left\{\frac{y_{ik}\theta_{ik}-b_k(\theta_{ik})}{a_k(\phi_k)}+c_k(y_{ik};\phi_k)\right\},\tag{1}$$

where θ_{ik} is the natural parameter of y_{ik}, ϕ_k is the dispersion parameter of the kth outcome variable, and a_k,b_k,c_k are known functions determined by the specific distribution of the kth outcome variable. Here the q outcome variables are allowed to have different distributions in the exponential-dispersion family; Table 1 provides details of some of the most common distributions in this family, including the Normal, Bernoulli, and Poisson distributions. The dispersion parameter ϕ_k can either be known or unknown. For example, ϕ_k for the Poisson distribution equals 1 which is known, but for the Gaussian distribution ϕ_k corresponds to the variance parameter which may be estimated from the data. Let $\boldsymbol{\phi}=(\phi_1,\ldots,\phi_q)^\top$ be the vector of the dispersion parameters, and denote $\boldsymbol{\phi}_u$ as a subvector of $\boldsymbol{\phi}$ consisting of all the unknown dispersion parameters. Without loss of generality, for each outcome variable, we apply the canonical link function $g_k=(b_k')^{-1}$, so that $\mathrm{E}(y_{ik})=b_k'(\theta_{ik})=g_k^{-1}(\theta_{ik})$, where b_k' denotes the derivative function of b_k.

Let $\mathbf{X}=(\tilde{\mathbf{x}}_1,\ldots,\tilde{\mathbf{x}}_p)=(\mathbf{x}_1,\ldots,\mathbf{x}_n)^\top\in\mathbb{R}^{n\times p}$ be the observed feature/predictor matrix, where the number p of features can be much larger than the sample size n. Also let $\mathbf{Z}=(\mathbf{z}_1,\ldots,\mathbf{z}_n)^\top\in\mathbb{R}^{n\times(p_z+1)}$ be consisting of a vector of 1s in its first column (to be corresponding to the intercept term) and the observed data from a few control variables in its subsequent p_z columns. The choice of control variables depends on the application, e.g., gender and age. Here it is understood that the number p_z of control variables is much smaller than the sample size n, which is the case in most real applications. The skeleton of our proposed approach is the familiar generalized linear model (GLM). Specifically, we model the natural parameters in (1) as

$$\theta_{ik}=o_{ik}+\mathbf{z}_i^\top\boldsymbol{\beta}_k+\mathbf{x}_i^\top\mathbf{c}_k,\tag{2}$$

where $(i,k)\in\Omega$ and the o_{ik}s are known offset terms. The offset terms commonly arise, for example, in the modeling of count data for adjusting the size of a population from which a count is drawn. The $\boldsymbol{\beta}_k$s are unknown coefficient vectors corresponding to the intercept and the control variables, and $\mathbf{c}_k$ are unknown coefficient vectors corresponding to the high-dimensional predictors. Let $\mathbf{C}=(\mathbf{c}_1,\ldots,\mathbf{c}_q)=(\tilde{\mathbf{c}}_1,\ldots,\tilde{\mathbf{c}}_p)^\top\in\mathbb{R}^{p\times q}$ be the coefficient matrix of the predictors and $\boldsymbol{\beta}=(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_q)=(\tilde{\boldsymbol{\beta}}_0,\ldots,\tilde{\boldsymbol{\beta}}_{p_z})^\top\in\mathbb{R}^{(p_z+1)\times q}$ be the coefficient matrix of the intercept and control variables, whose first row, $\tilde{\boldsymbol{\beta}}_0$, gives the intercept vector.

Without any additional assumptions on the parameters, the above model is over-parameterized when p or q are comparable or much larger than n. Moreover, when the independence of the y_{ik}s is assumed, the model reduces to a set of univariate GLM analysis and thus does not possess any multivariate flavor. The key here is how to induce and take advantage of the dependence among the outcomes. We argue that the merit of the formulations in (1) and (2) lies in imposing some suitable low-dimensional structures on $\mathbf{C}$. In particular, here we consider the case where $\mathbf{C}$ is a *reduced-rank or low-rank matrix*. The low-rank assumption implies that the outcomes $\mathbf{Y}$ are dependent on the predictors through a few latent variables. To see this, assume the rank of $\mathbf{C}$ is r, which can be much smaller than both p and q. Then $\mathbf{C}$ can be decomposed as $\mathbf{C}=\mathbf{A}\mathbf{B}^\top$, for some $\mathbf{A}\in\mathbb{R}^{p\times r}$ and $\mathbf{B}\in\mathbb{R}^{q\times r}$. From this decomposition, all the outcomes $\mathbf{Y}$ are linked to the predictors $\mathbf{X}$ only through a few latent directions $\mathbf{XA}$, which are some unknown linear combinations of the original predictors and are what we refer to as "latent variables" or "latent factors". Because these latent variables are shared among all the responses, a reduced-rank model indeed possesses a genuine multivariate flavor. Another useful perspective is to view the proposed model as a "supervised" factor analysis (FA); the main difference is that the latent factors in the proposed model are assumed to live in the subspace spanned by the predictors $\mathbf{X}$, rather than left unsupervised in FA. The proposed approach also connects to the sufficient dimension reduction (SDR) [15,29–31], since in mRRR the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ is the same as that of $\mathbf{Y}$ given $\mathbf{XA}$, albeit that this SDR structure is embedded in a parametric GLM framework. With this information-sharing mechanism enabled by the low-rank structure, we then assume that the outcomes y_{ik} are conditionally independent given the predictors. This appealing latent model setup facilities model interpretation, enables dimension reduction for handling high dimensional data, and more importantly, in an implicit way it induces dependency among the outcomes.

We term the above model, i.e., (1), (2) together with the low-rank assumption on $\mathbf{C}$, as a *mixed-response reduced-rank regression* (mRRR) model. The mRRR is a rather "simple" model, in the sense that the dependency among the outcomes is not explicitly modeled. Indeed, in this research, we do not intend to incorporate certain fully-parameterized association matrix to explicitly characterize the dependency among the outcomes or specify a comprehensive joint distribution for them. Such proposals would not be easily applicable or generalizable in the simultaneous presence of outcome heterogeneity, high dimensionality and incomplete data. Our focus is to show that the reduced-rank methodology has great potential in handling such complex data structures. We will also discuss later that the mRRR model can be further generalized in various aspects, such as the incorporation of variable selection, outlier detection and/or robust estimation.

We conduct mRRR analysis via the following regularized estimation approach:

$$\min_{\mathbf{C},\boldsymbol{\beta},\boldsymbol{\phi}_u} \left[F(\mathbf{C},\boldsymbol{\beta},\boldsymbol{\phi}_u) \equiv -\sum_{(i,k)\in\Omega} \ell_k(\mathbf{c}_k,\boldsymbol{\beta}_k,\phi_k;\mathbf{x}_i,\mathbf{z}_i,y_{ik}) + \sum_{h=1}^{p\wedge q} \rho\{d_h(\mathbf{C});\lambda\} \right], \tag{3}$$

where

$$\ell_k(\mathbf{c}_k,\boldsymbol{\beta}_k,\phi_k;\mathbf{x}_i,\mathbf{z}_i,y_{ik}) = \frac{y_{ik}(o_{ik}+\mathbf{z}_i^\top\boldsymbol{\beta}_k+\mathbf{x}_i^\top\mathbf{c}_k) - b_k(o_{ik}+\mathbf{z}_i^\top\boldsymbol{\beta}_k+\mathbf{x}_i^\top\mathbf{c}_k)}{a_k(\phi_k)} + c_k(y_{ik},\phi_k) \tag{4}$$

is the log-likelihood function, $d_h(\mathbf{C})$ denotes the hth largest singular value of $\mathbf{C}$, $\rho(\cdot;\lambda)$ is a penalty function with tuning parameter λ for inducing the sparsity of the singular values and hence reducing the rank. The intercept term and the effects of control variables are always included and not penalized. Some popular choices for the penalty include the convex nuclear norm penalty, i.e.,

$$\sum_{h=1}^{p\wedge q} \rho\{d_h(\mathbf{C});\lambda\} = \lambda \sum_{h=1}^{p\wedge q} d_h(\mathbf{C}) = \lambda\|\mathbf{C}\|_*, \tag{5}$$

where $\|\cdot\|_*$ denotes the nuclear norm, and the nonconvex rank penalty, i.e.,

$$\sum_{h=1}^{p\wedge q} \rho\{d_h(\mathbf{C});\lambda\} = \lambda \sum_{h=1}^{p\wedge q} \mathbf{1}\{d_h(\mathbf{C})>0\} = \lambda r(\mathbf{C}), \tag{6}$$

where $\mathbf{1}(\cdot)$ is the indicator function and r denotes the rank of the enclosed matrix. It can be readily shown that possible solutions produced by the rank penalized approach, i.e., (3) with (6), can also be obtained via the rank-constrained estimation as follows [10],

$$\min_{\mathbf{C},\boldsymbol{\beta},\boldsymbol{\phi}_u} \left\{ -\sum_{(i,k)\in\Omega} \ell_k(\mathbf{c}_k,\boldsymbol{\beta}_k,\phi_k;\mathbf{x}_i,\mathbf{z}_i,y_{ik}) \right\} \quad \text{such that } r(\mathbf{C}) \le r, \tag{7}$$

for $r \in \{1,\ldots,p\wedge q\}$. In the current work we mainly focus on the rank-penalized version of mRRR.

When $\mathbf{X}$ is taken as the $n \times n$ identity matrix, mRRR specializes to a mixed principal component analysis (PCA) with incomplete data. When the responses are all from the same type of distribution, the model further simplifies to a generalized PCA [13] or a generalized matrix completion [7]. When all the outcomes are Gaussian, mRRR reduces to the regular RRR, in which the dispersion parameter can be treated as nuisance.

3. Non-asymptotic analysis

Building upon a rich literature on low rank models, e.g., [4,27,28,51], we study the performance of mRRR in a general context of supervised learning with mixed and incomplete outcomes. Denote $s = |\Omega|$, i.e., the number of observed entries in $\mathbf{Y}$. For simplicity we omit the fixed and known offset terms o_{ik}s in (2), so the model for the natural parameters becomes

$$\theta_{ik} = \mathbf{x}_i^\top\mathbf{c}_k + \mathbf{z}_i^\top\boldsymbol{\beta}_k,$$

or in its matrix form, $\boldsymbol{\Theta} = \mathbf{Z}\boldsymbol{\beta} + \mathbf{XC}$. We consider fixed design. Let $\mathbf{A} = (\mathbf{Z},\mathbf{X}) \in \mathbb{R}^{n\times(p_z+p+1)}$, and let $\mathcal{P}_{\mathbf{A}} = \mathbf{A}(\mathbf{A}^\top\mathbf{A})^-\mathbf{A}^\top$ be the projection matrix onto the column space of $\mathbf{A}$, where $(\cdot)^-$ denotes the Moore–Penrose inverse. We also use $\mathcal{P}_{\mathbf{A}}^\perp$ to denote the projection matrix onto the orthogonal complement of the column space of $\mathbf{A}$.

We consider $\mathbf{C}$ and $\boldsymbol{\beta}$ in a bounded parameter space,

$$\mathcal{C} = \{\boldsymbol{\beta} \in \mathbb{R}^{p_z\times q}, \mathbf{C} \in \mathbb{R}^{(p+1)\times q}, |\theta_{ik}| \le K, i \in \{1,\ldots,n\}, k \in \{1,\ldots,q\}\}.$$

Here $(\mathbf{C}^*,\boldsymbol{\beta}^*) \in \mathcal{C}$ is to denote the parameters of the underlying true model, where $\mathbf{C}^*$ is possibly of low rank and $\boldsymbol{\beta}^*$ is of full column rank. Also let $r^* = r(\mathbf{C}^*)$, and $\boldsymbol{\Theta}^* = \mathbf{Z}\boldsymbol{\beta}^* + \mathbf{XC}^*$.

Regarding the model structure, in this work we assume that the dispersion parameters are known and focus mainly on the estimation of the natural parameters. For simplicity and without loss of generality, we then let $a(\phi_k) = 1$ and consequently $f(y_{ik}; \theta_{ik}) = \exp\{y_{ik}\theta_{ik} - b(\theta_{ik}) + c(y_{ik}; \phi_k)\}$. The log-likelihood becomes

$$L(\mathbf{C}, \boldsymbol{\beta}; \mathbf{X}, \mathbf{Z}, \widetilde{\mathbf{Y}}) = \sum_{(i,k)\in\Omega} \ell_k(\mathbf{c}_k, \boldsymbol{\beta}_k; \mathbf{x}_i, \mathbf{z}_i, y_{ik}) = \langle \widetilde{\mathbf{Y}}, \boldsymbol{\Theta}\rangle_F - \langle \mathbf{b}(\boldsymbol{\Theta}), \mathbf{1}_{\Omega}^{n\times q}\rangle_F + \text{const},$$

where $\langle\cdot\rangle_F$ denotes the Frobenius inner product, $\mathbf{1}_{\Omega}^{n\times q}$ is a $n \times q$ matrix with those entries with indices in Ω being 1 and other entries being 0, and $\mathbf{b}(\boldsymbol{\Theta}) = (b_1(\boldsymbol{\theta}_1), \ldots, b_q(\boldsymbol{\theta}_q))$, with $b_k(\boldsymbol{\theta}_k) = (b_k(\theta_{1k}), \ldots, b_k(\theta_{nk}))^\top$ for $k \in \{1, \ldots, q\}$.

Condition 1. *For each $k \in \{1, \ldots, q\}$, b_k is a continuously differentiable, real-valued and strictly convex function defined on a closed convex set. Also, for some constants $\overline{\gamma}, \underline{\gamma} > 0$,*

$$\max_{i\in\{1,\ldots,n\},\, k\in\{1,\ldots,q\}} \sup_{\boldsymbol{\beta},\mathbf{C}\in\mathcal{C}} |b_k''(\theta_{ik})| \le \overline{\gamma} \quad \textit{and} \quad \min_{i\in\{1,\ldots,n\},\, k\in\{1,\ldots,q\}} \inf_{\boldsymbol{\beta},\mathbf{C}\in\mathcal{C}} |b_k''(\theta_{ik})| \ge \underline{\gamma}.$$

Remark 1. If b_k is smooth enough, a simple derivation of the density shows that its successive derivatives can be used to determine the distribution moments. When b_k is twice differentiable, $\mathrm{E}(y_{ik}|\theta_{ik}^*) = b_k'(\theta_{ik}^*)$ and $\mathrm{var}(y_{ik}|\theta_{ik}^*) = b_k''(\theta_{ik}^*)$ hold for $(i, k) \in \Omega$.

The error matrix is defined as the difference between the response matrix and its expectation, $\mathbf{E} = \widetilde{\mathbf{Y}} - P_\Omega\{\boldsymbol{\mu}(\boldsymbol{\Theta}^*)\}$. To cover a wide range of outcomes, we assume the entries of $\mathbf{E} = (e_{ik})$ follow a sub-exponential distribution.

Condition 2. *Assume that $\mathbf{E}$ has independent entries and for some constant $\sigma_E > 0$,*

$$\max_{i\in\{1,\ldots,n\},\, k\in\{1,\ldots,q\}} \mathrm{E}\{\exp(|e_{ik}|/\sigma_E)\} \le e.$$

Here, for simplicity σ_E is the same for all the responses.

We regard the incompleteness of the data as due to sampling. Following Klopp [26] and Lafond [28], we impose conditions on the boundedness of the sampling probabilities and the number of entries.

Condition 3. *Let $\pi_{ik} = \Pr\{(i, k) \in \Omega\}$ for all $i \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, q\}$. Assume that for some constants $\mu, \nu \ge 1$,*

$$\min_{i\in\{1,\ldots,n\},\, k\in\{1,\ldots,q\}} \pi_{ik} \ge \frac{1}{\mu n q}, \tag{8}$$

$$\max_{i\in\{1,\ldots,n\},\, k\in\{1,\ldots,q\}} \left(\sum_i \pi_{ik}, \sum_k \pi_{ik}\right) \le \frac{\nu}{n \wedge q}. \tag{9}$$

The condition in (8) bounds the lowest sampling probability of all entries and (9) ensures that no row nor column should be sampled far more frequently than the others. When $\mu = \nu = 1$, the condition corresponds to the special case of uniform sampling. The estimation may fail when the number of observed entries is too small or the noise in the observed outcomes is too large. The next condition then imposes a lower bound on the number of observed entries which depends on the noise level measured by σ_E^2.

Condition 4.

$$s = |\Omega| > \frac{2}{\nu} \ln\{r(\mathbf{A}) + q\}(n \wedge q) \max\left\{\frac{\sigma_E^2}{\overline{\gamma}} \ln^2\left(\sigma_E \sqrt{\frac{n\wedge q}{\underline{\gamma}}}\right), \frac{1}{9}\right\}.$$

The above conditions are reasonable in many practical scenarios of mixed response regression. In particular, Conditions 1 and 2 cover mixed responses from several commonly used distributions including the ones from the natural exponential family, such as the Bernoulli, Poisson, Normal distribution with fixed variance, Gamma distribution with fixed shape parameter, among others.

Consider the rank penalized mRRR estimator,

$$(\widehat{\mathbf{C}}, \widehat{\boldsymbol{\beta}}) = \arg\min_{\mathbf{C},\boldsymbol{\beta}\in\mathcal{C}} -\frac{1}{s} L(\mathbf{C}, \boldsymbol{\beta}; \mathbf{X}, \mathbf{Z}, \widetilde{\mathbf{Y}}) + \lambda r(\mathbf{C}). \tag{10}$$

We establish an oracle inequality for $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F^2$, where $\widehat{\boldsymbol{\Theta}} = \mathbf{X}\widehat{\mathbf{C}} + \mathbf{Z}\widehat{\boldsymbol{\beta}}$.

Theorem 1. *Suppose Conditions 1–4 hold. Choose*

$$\lambda = 16\mu\nu\left(\alpha^2 e^2 \underline{\gamma} + c_E^2 \frac{\overline{\gamma}}{\underline{\gamma}}\right) \frac{\ln\{r(\mathbf{A}) + q\}}{s} (n \vee q),$$

where α and c_E are constants defined in Lemma A.2. Then, with probability of at least $1 - 3\{r(\mathbf{A}) + q\}^{-1}$, we have

$$\frac{\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F^2}{nq} \lesssim \max\left[\mu e K^2 \sqrt{\frac{\ln\{r(\mathbf{A}) + q\}}{s}}, \mu^2 \nu \left(\alpha^2 e^2 + c_E^2 \frac{\overline{\gamma}}{\underline{\gamma}^2}\right)\{r(\mathbf{C}^*) + r(\mathbf{Z}\boldsymbol{\beta}^*)\} \frac{\ln\{r(\mathbf{A}) + q\}}{s} (n \vee q)\right],$$

where $\lesssim$ means that the inequality holds up to some multiplicative numerical constants.

Theorem 1 extends the existing results on high-dimensional reduced-rank regression. In general case, the second term dominates when n is large, and thus the mRRR estimator achieves

$$\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F^2 \lesssim \{r(\mathbf{C}^*) + p_z + 1\}(n \vee q) \ln\{r(\mathbf{A}) + q\}\, nq/s.$$

Comparing to the results in [4] that the reduced-rank regression estimator achieves $\|\mathbf{X}\widehat{\mathbf{C}} - \mathbf{X}\mathbf{C}^*\| \lesssim r(\mathbf{C}^*)(n \vee q)$ under sub-Gaussian error, the term $p_z + 1$ comes from the inclusion of the unpenalized control variables and intercept, the term $(nq)/s$ is due to the incomplete data, and the extra $\ln\{r(\mathbf{A}) + q\}$ term arises from the sub-exponential error structure.

The nuclear-norm penalized estimator using (5) can achieve the same error rate as the rank penalized mRRR given in Theorem 1, albeit under additional conditions on the design matrix [4,43]. See, for example, Theorem 12 in [4], which states that under full-rank design, nuclear-norm penalized regression can achieve the same error bound as its rank penalized counterpart. We omit the details.

4. Computation

We first briefly review a generic matrix approximation problem, viz.

$$\min_{\boldsymbol{\Gamma} \in \mathbb{R}^{n \times q}} \frac{1}{2} \|\mathbf{Y} - \boldsymbol{\Gamma}\|_F^2 + \sum_{h=1}^{n \wedge q} \rho\{d_h(\boldsymbol{\Gamma}); \lambda\}, \tag{11}$$

where $\|\cdot\|_F$ denotes the Frobenius norm and ρ is a sparsity-inducing penalty. Let $\mathbf{Y} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ be the singular value decomposition (SVD) of $\mathbf{Y}$, so that $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_{n \wedge q}$, $\mathbf{V}^\top\mathbf{V} = \mathbf{I}_{n \wedge q}$ where $\mathbf{I}_{n \wedge q}$ is the $(n \wedge q) \times (n \wedge q)$ identity matrix, and $\mathbf{D}$ is a diagonal matrix with the nonzero singular values on its diagonal in descending order. For a given ρ, the solution of (11), $\widehat{\boldsymbol{\Gamma}}$, is obtained from singular value thresholding by $\mathbb{T}^d(\cdot; \lambda)$, i.e.,

$$\widehat{\boldsymbol{\Gamma}} = \mathbb{T}^d(\mathbf{Y}; \lambda) \equiv \mathbf{U}\mathbb{T}(\mathbf{D}; \lambda)\mathbf{V}^\top, \tag{12}$$

where $\mathbb{T}(\cdot; \lambda)$ is an element-wise thresholding function associated with the penalty ρ. For example, the ℓ_1 penalty is associated with the soft-thresholding operator, i.e., $\mathbb{T}(t; \lambda) = \text{sign}(t)(|t| - \lambda)_+$, and the ℓ_0 penalty is associated with the hard-thresholding operator, i.e., $\mathbb{T}(t; \lambda) = t\mathbf{1}(|t| > \sqrt{2\lambda})$. Consequently, (11) with the nuclear norm penalization is solved by singular value soft-thresholding, while (11) with the rank penalization is solved by singular value hard-thresholding [6,10,43].

Utilizing the above results, we present an iterative singular value thresholding algorithm to solve the mRRR problem in (3). To save space, its derivation is given in the Appendices A and B. For the ease of presentation, we omit the o_{ik} term in (2) as their inclusion adds no difficulty in computation. Define $\boldsymbol{\Phi} = \text{diag}\{a_1(\phi_1), \ldots, a_q(\phi_q)\}$, a diagonal matrix with $a_k(\phi_k)$s on its diagonal. Define

$$\boldsymbol{\mu}(\mathbf{c}_k, \boldsymbol{\beta}_k) = (g_k^{-1}(\mathbf{x}_1^\top\mathbf{c}_k + \mathbf{z}_1^\top\boldsymbol{\beta}_k), \ldots, g_k^{-1}(\mathbf{x}_n^\top\mathbf{c}_k + \mathbf{z}_n^\top\boldsymbol{\beta}_k))^\top \in \mathbb{R}^n, \quad \boldsymbol{\mu}(\mathbf{C}, \boldsymbol{\beta}) = (\boldsymbol{\mu}(\mathbf{c}_1, \boldsymbol{\beta}_1), \ldots, \boldsymbol{\mu}(\mathbf{c}_q, \boldsymbol{\beta}_q)) \in \mathbb{R}^{n \times q}.$$

The proposed algorithm is given in Algorithm 1.

Algorithm 1 Mixed-response reduced-rank regression algorithm (mRRR)

```
Initialize $\mathbf{C}^{(0)}$, $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\phi}_u^{(0)}$. Set $t \leftarrow 0$.
repeat
    (a) C-step: $\mathbf{C}^{(t+1)} = \mathbb{T}^d[\mathbf{C}^{(t)} + \mathbf{X}^\top P_\Omega\{\mathbf{Y} - \boldsymbol{\mu}(\mathbf{C}^{(t)}, \boldsymbol{\beta}^{(t)})\}\boldsymbol{\Phi}^{(t)-1}; \lambda]$,
    (b) β-step: $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \mathbf{Z}^\top P_\Omega\{\mathbf{Y} - \boldsymbol{\mu}(\mathbf{C}^{(t+1)}, \boldsymbol{\beta}^{(t)})\}\boldsymbol{\Phi}^{(t)-1}$,
    (c) φ_u-step: $\boldsymbol{\phi}_u^{(t+1)} = \arg\max_{\boldsymbol{\phi}_u} \sum_{(i,k)\in\Omega} \ell_k(\mathbf{c}_k^{(t+1)}, \boldsymbol{\beta}_k^{(t+1)}, \phi_k; \mathbf{x}_i, \mathbf{z}_i, y_{ik})$,
    $t \leftarrow t + 1$.
until convergence,
    e.g., $|F(\mathbf{C}^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\phi}_u^{(t)}) - F(\mathbf{C}^{(t+1)}, \boldsymbol{\beta}^{(t+1)}, \boldsymbol{\phi}_u^{(t+1)})| / |F(\mathbf{C}^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\phi}_u^{(t)})| \le \epsilon$, e.g., $\epsilon = 10^{-6}$.
return $\widehat{\mathbf{C}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}_u$.
```

Several remarks are in order. The **C**-step is solved by singular value thresholding. As in practice the desired rank is usually much smaller than both p and q, the computation cost of performing SVD can be well controlled. When updating the unknown dispersion parameters, the problem is separable in each ϕ_k, and can be handled by standard optimization methods such as Newton–Raphson. Besides, for some common distributions, e.g., Gaussian, the optimizer of a dispersion parameter admits an explicit form. Algorithm 1 can be readily modified to optimize the rank-constrained criterion in (7), for which the $\mathbf{C}$-step is replaced with setting $\mathbf{C}^{(t+1)}$ as the rank-r approximation of $\mathbf{C}^{(t)} + \mathbf{X}^\top P_\Omega\{\mathbf{Y} - \boldsymbol{\mu}(\mathbf{C}^{(t)}, \boldsymbol{\beta}^{(t)})\}\boldsymbol{\Phi}^{(t)-1}$.

We have the following results regarding the convergence properties of Algorithm 1. Define

$$\mathbf{W}(\mathbf{c}_k, \boldsymbol{\beta}_k, \phi_k) = \operatorname{diag}\{|b_k''(\mathbf{x}_i^\top \mathbf{c}_k + \mathbf{z}_i^\top \boldsymbol{\beta}_k)|/a_k(\phi_k) : i \in \{1, \ldots, n\}\}, \quad \gamma_1 = \max_{k\in\{1,\ldots,q\}} \sup_{(\boldsymbol{\xi}_k,\boldsymbol{\zeta}_k,\delta_k)\in\mathcal{A}_k} \|\mathbf{X}^\top \mathbf{W}(\boldsymbol{\xi}_k, \boldsymbol{\zeta}_k, \delta_k)\mathbf{X}\|_2,$$

where $\mathcal{A}_k = \{(a\mathbf{c}_k^{(t)} + (1-a)\mathbf{c}_k^{(t+1)}, \boldsymbol{\beta}_k^{(t)}, \phi_k^{(t)}) : a \in (0, 1), t \in \{1, 2, \ldots\}\}$, and

$$\gamma_2 = \max_{k\in\{1,\ldots,q\}} \sup_{(\boldsymbol{\xi}_k,\boldsymbol{\zeta}_k,\delta_k)\in\mathcal{B}_k} \|\mathbf{Z}^\top \mathbf{W}(\boldsymbol{\xi}_k, \boldsymbol{\zeta}_k, \delta_k)\mathbf{Z}\|_2.$$

where $\mathcal{B}_k = \{(\mathbf{c}_k^{(t+1)}, a\boldsymbol{\beta}_k^{(t)} + (1-a)\boldsymbol{\beta}_k^{(t+1)}, \phi_k^{(t)}) : a \in (0, 1), t \in \{1, 2, \ldots\}\}$. Here $\|\cdot\|_2$ denotes the spectral norm.

Theorem 2. *The sequence* $\{\mathbf{C}^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\phi}_u^{(t)}\}$ *produced by Algorithm 1 satisfies,*

$$F(\mathbf{C}^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\phi}_u^{(t)}) - F(\mathbf{C}^{(t+1)}, \boldsymbol{\beta}^{(t+1)}, \boldsymbol{\phi}_u^{(t+1)}) \geq \frac{1-\gamma_1}{2}\|\mathbf{C}^{(t+1)} - \mathbf{C}^{(t)}\|_F^2 + \frac{2-\gamma_2}{2}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_F^2.$$

Theorem 2 shows that as long as $\gamma_1 \leq 1, \gamma_2 \leq 2$, the monotone descending of the objective function is guaranteed. Following She [42], we achieve this by properly scaling $\mathbf{X}$ and $\mathbf{Z}$. The upper bound of γ_1 can be determined for several common distributions. For Gaussian responses, $b_k''(x) = 1$ for any x, and $a_k(\phi_k) = \sigma_k^2$, where σ_k^2 is the variance parameter; therefore, $\gamma_1 \leq \|\mathbf{X}\|_2^2/\min(\sigma_k^2)$, where the minimum is over all the Gaussian responses. It then suffices to scale $\mathbf{X}$ by some scaling factor $\kappa_1^* \geq \|\mathbf{X}\|_2/\min(\sigma_k)$, so that after scaling it comes that $\gamma_1 \leq 1$. In practice, σ_k can be replaced by some conservative initial estimator, e.g., from fitting linear regression of Gaussian response alone. For binary responses, $b_k''(x) = e^x/(1+e^x)^2 \leq 1/4$ for any x and $a_k(\phi_k) = 1$, so that $\gamma_1 \leq \|\mathbf{X}\|_2^2/4$. Then it suffices to set $\kappa_1^* = \|\mathbf{X}\|_2/2$. For Poisson responses, $b_k''(x) = e^x$, so the upper bound of γ_1 does not have a simple form. In practice when an explicit bound is not available, we empirically set a large enough scaling factor to ensure the descending of the $\mathbf{C}$-step, with the expense of reduced convergence speed. In the mixed response setting, the scale factor can be chosen as the maximum of these quantities for different types of distributions. Similarly, $2 - \gamma_2 \geq 0$ can be achieved by scaling $\mathbf{Z}$.

We stress that γ_1, γ_2 and the aforementioned choice of the scaling factor do not depend on the tuning parameter and the penalty form in the objective function, and an estimator from the scaled model can be simply scaled back to give the solution of the original problem. Although Theorem 2 does not guarantee the convergence of $\{\mathbf{C}^{(t)}\}$ in general, in practice we always observe a unique limit point, and the algorithm is efficient and stable.

To initialize Algorithm 1, we set $\mathbf{C}^{(0)} = \mathbf{0}$ and obtain $\boldsymbol{\beta}^{(0)}, \boldsymbol{\phi}_u^{(0)}$ from univariate GLMs. When the model is fitted for a sequence of λ values, the warm start strategy is adopted, i.e., using the solution from previous fit as the initial value for the next λ value. We use K-fold cross validation [46] to choose the optimal λ and hence the optimal solution, based on the predictive performance of the models. The implementation is available in the R package rrpack.

5. Simulation

5.1. Simulation setups

We consider several simulation models. Model 1 is a low-dimensional example. We set $n = 100, p = 15, q = 20$, and $r = 2$. Among the $q = 20$ responses, $q_1 = 8$ of them are generated from Gaussian, $q_2 = 10$ from Bernoulli, and $q_3 = 2$ from Poisson. The predictor matrix $\mathbf{X}$ is constructed by generating its entries as independent and identically distributed (iid) random samples from the standard normal distribution $\mathcal{N}(0, 1)$. The coefficient matrix $\mathbf{C}$ is generated as $\mathbf{C} = \mathbf{A}\mathbf{B}^\top$, where $\mathbf{A} \in \mathbb{R}^{p\times r}$ is an orthogonal matrix from the QR decomposition of a random $p \times r$ matrix filled with $\mathcal{N}(0, 1)$ entries, and all entries in $\mathbf{B} \in \mathbb{R}^{q\times r}$ are iid samples from $\mathcal{U}(-1, 1)$. We let $\mathbf{Z} = \mathbf{1}_n$, and set the corresponding coefficient matrix, i.e., the intercept vector, as $\boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}_0 = 0.5\mathbf{1}_q$. The natural parameter matrix is then constructed as $\boldsymbol{\Theta} = (\theta_{ik})_{n\times q} = \mathbf{Z}\boldsymbol{\beta} + \mathbf{X}\mathbf{C}$. In this example, all the Gaussian responses are set to have the same dispersion parameter, i.e., $y_{ik} \sim \mathcal{N}(\theta_{ik}, 1)$ for all $i \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, q_1\}$. The response matrix $\mathbf{Y}$ is then generated from model (1).

Model 2 is a high-dimensional setup. We set $n = 200, p = 1000, q = 20$ and $r = 2$. The responses still consist of $q_1 = 8$ Gaussian, $q_2 = 10$ Bernoulli, and $q_3 = 2$ Poisson variables. The matrix $\mathbf{X}$ is generated as $\mathbf{X}_1\mathbf{X}_2^\top$, where $\mathbf{X}_1 \in \mathbb{R}^{n\times 10}$, $\mathbf{X}_2 \in \mathbb{R}^{p\times 10}$, and all entries of $\mathbf{X}_1$ and $\mathbf{X}_2$ are iid $\mathcal{N}(0, 1)$; the entire matrix is then scaled so that $\|\mathbf{X}\|_F^2/(np) = 1$. The $\mathbf{C}, \mathbf{Z}, \boldsymbol{\beta}$ and $\boldsymbol{\Theta}$ are generated in the same way as in Model 1. The Gaussian responses are set to have different dispersion parameters, i.e., $y_{ik} \sim \mathcal{N}(\theta_{ik}, k/q_1)$, for $i \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, q\}$.

In each simulation run, once the full data $(\mathbf{X}, \mathbf{Y})$ are generated, we randomly choose M% of entries in $\mathbf{Y}$ and set them as missing values, where $M \in \{0, 10, 20\}$. The simulation experiments are replicated 100 times under each model setting.

5.2. Methods and evaluation metrics

We consider several realistic modeling strategies for mixed responses/outcomes and incomplete data of possibly high dimension. A simple approach is to model the responses marginally and separately, by fitting a univariate generalized linear model for each response (uGLM). In each univariate regression, only complete data pairs $(y_{ik}, \mathbf{x}_i)$ are used, and when dealing with high dimension, we have used its penalized version with the elastic net penalty [55] implemented in the `R` package `glmnet`. Another approach is to fit vector generalized reduced-rank regression (gRRR) [42,50] for each type of responses separately. The gRRR can be viewed as a special case of mRRR, so our proposed algorithm still applies and also enables gRRR to handle incomplete responses. To focus on the main idea and ease the presentation, however, we only separate the Gaussian and non-Gaussian variables, e.g., binary and count data are modeled together using mRRR. With the proposed mRRR, we are able to model all the mixed outcomes simultaneously. Both the rank and the nuclear-norm penalized gRRR and mRRR are considered, and the corresponding methods are denoted as gRRR.r/gRRR.n and mRRR.r/mRRR.n, respectively. As a benchmark, we also include an oracle approach by fitting mRRR with complete data (ORE); similarly, the methods with different penalties are denoted as ORE.r and ORE.n.

For each reduced-rank method, we use 5-fold cross-validation for tuning parameter selection. To be specific, we split the non-missing entries in $\mathbf{Y}$ to five folds, and each time use one fold as testing set (Ω_{te}) and the others as training set (Ω_{tr}). We apply each method to obtain its solution path using the training set, and evaluate the models along the path by a predictive deviance measure using the testing set, i.e., $-2\sum_{(i,k)\in\Omega_{te}} \ell_k(\hat{\mathbf{c}}_k, \hat{\phi}_k; \tilde{\mathbf{x}}_i, y_{ik})$. The optimal tuning parameter and hence the optimal model is selected as the one with the smallest cross validation error.

Let $\boldsymbol{\Theta}_1$, $\boldsymbol{\Theta}_2$ be the sub-matrices of $\boldsymbol{\Theta}$ corresponding to the Gaussian and non-Gaussian outcomes, respectively. The estimation of $\boldsymbol{\Theta}$ is evaluated by

$$\mathrm{Er}_g(\widehat{\boldsymbol{\Theta}}) = \|\widehat{\boldsymbol{\Theta}}_1 - \boldsymbol{\Theta}_1\|_F^2/(nq_1), \quad \mathrm{Er}_{ng}(\widehat{\boldsymbol{\Theta}}) = \|\widehat{\boldsymbol{\Theta}}_2 - \boldsymbol{\Theta}_2\|_F^2/\{n(q-q_1)\},$$

and the combined estimation error $\mathrm{Er}(\widehat{\boldsymbol{\Theta}}) = \|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_F^2/(nq)$. Moreover, the estimation error of the dispersion parameters is computed as

$$\mathrm{Er}(\widehat{\boldsymbol{\phi}}) = \left\{\sum_{k=1}^{q_1} (\hat{\phi}_k - \phi_k)^2\right\}/q_1.$$

To evaluate the rank estimation performance, we present the average selected rank (Rank), and the ratio between the estimated nuclear norm and the truth, i.e., $\mathrm{Ratio}_* = \|\widehat{\mathbf{XC}}\|_*/\|\mathbf{XC}\|_*$.

5.3. Simulation results

The simulation results of Models 1 and 2 are reported in Tables 2–3, respectively. In Figs. 1–2, we also display the boxplots of combined estimation error. As expected, the uGLM strategy performs the worst among all methods in every category. This is because such a marginal method fails to exploit the potential correlation among the outcomes. In a reduced-rank model, all the predictors may contribute to the prediction of the outcomes, which is quite different from the sparse model assumption. As such, the sparse and shrinkage estimation adopted in uGLM may not be able to mimic the desired model structure in these examples.

Comparing the two joint estimation approaches, mRRR performs substantially better than gRRR which models Gaussian and non-Gaussian outcomes separately. We have also tried modeling separately each type of distributions, and the results are even worse than gRRR and hence are not reported. The reason is that gRRR is only able to partially capture the correlation within each type of outcomes, while mRRR fully captures the latent dependency among all the outcomes. With incomplete data, the performance of mRRR, gRRR and uGLM all becomes worse as the proportion of missing increases. Nevertheless, mRRR still performs comparably well to ORE.

Comparing the rank penalized mRRR and the nuclear norm penalized mRRR, the performance of the latter is generally worse than that of the former. In particular, using nuclear norm penalty tends to select a larger rank, while the overall nuclear norm gets much heavier shrinkage. These findings agree with existing studies [10,51].

6. Application in longitudinal studies of aging

The Longitudinal Studies of Aging (LSOAs) is a collaborative project conducted by the US National Center for Health Statistics and the National Institute on Aging [45]. A national representative sample of several thousands of subjects who were at or over 70 years of age were interviewed and followed, and their health, functional status, living arrangements, and health services utilization were measured as they moved into and through their oldest ages. It is of interest to examine the changes and the associations between their current and future health status. Therefore, we consider a multivariate regression setup, by jointly regressing various health measures collected during the period 1999–2000 on the records collected during the period 1997–1998 from the same set of subjects. There are in total $n = 3988$ subjects who participated in the studies in both periods.

There are $q = 44$ outcome variables covering a wide range of assessments of health conditions. Specifically, three self-rated health measures, including overall health status, memory status and depression status, can be regarded as continuous

Table 2
Simulation: results of Model 1. Reported are the average values of various performance measures over replicated simulation experiments, with their standard deviations reported in parentheses. To improve presentation, $\mathrm{Er}_g(\widehat{\Theta})$, $\mathrm{Er}_{ng}(\widehat{\Theta})$ and $\mathrm{Er}(\widehat{\boldsymbol{\phi}})$ are scaled by multiplying 10^2.

M%		ORE.r	mRRR.r	gRRR.r	ORE.n	mRRR.n	gRRR.n	uGLM
0%	$\mathrm{Er}_g(\widehat{\Theta})$	5.2 (1.2)	5.2 (1.2)	7.1 (2.4)	11.7 (1.7)	11.7 (1.7)	13.5 (1.9)	33.8 (3.9)
	$\mathrm{Er}_{ng}(\widehat{\Theta})$	17.1 (5.2)	17.1 (5.2)	24.9 (8.1)	27.1 (4.3)	27.1 (4.3)	43.2 (6.8)	82.3 (11.1)
	$\mathrm{Er}(\widehat{\boldsymbol{\phi}})$	2.2 (2.7)	2.2 (2.7)	2.4 (3.5)	4.1 (4.0)	4.1 (4.0)	4.3 (4.1)	4.4 (5.2)
	Rank	2.0 (0.0)	2.0 (0.0)	1.9 (0.1)/2.0 (0.1)	6.6 (0.5)	6.6 (0.5)	4.0 (0.5)/5.5 (0.5)	–
	Ratio_*	1.1 (0.0)	1.1 (0.0)	1.2 (0.1)	0.9 (0.1)	0.9 (0.1)	1.2 (0.1)	1.0 (0.1)
10%	$\mathrm{Er}_g(\widehat{\Theta})$	5.3 (1.2)	5.9 (1.3)	9.3 (4.0)	11.8 (1.5)	14.0 (1.8)	15.7 (2.0)	36.2 (4.5)
	$\mathrm{Er}_{ng}(\widehat{\Theta})$	18.9 (5.6)	21.4 (6.5)	30.9 (8.1)	26.8 (4.0)	31.4 (5.0)	51.1 (8.1)	81.2 (9.1)
	$\mathrm{Er}(\widehat{\boldsymbol{\phi}})$	2.3 (3.1)	4.2 (4.6)	5.0 (6.8)	4.1 (4.2)	8.2 (6.3)	8.3 (6.4)	6.2 (8.0)
	Rank	2.0 (0.1)	2.0 (0.1)	2.0 (0.1)/2.0 (0.2)	6.7 (0.5)	6.8 (0.4)	4.0 (0.5)/5.5 (0.5)	–
	Ratio_*	1.1 (0.0)	1.1 (0.0)	1.2 (0.1)	0.9 (0.1)	0.9 (0.1)	1.2 (0.1)	1.0 (0.1)
20%	$\mathrm{Er}_g(\widehat{\Theta})$	5.2 (1.2)	7.8 (4.4)	13.6 (7.0)	12.3 (1.7)	17.2 (2.5)	19.2 (2.7)	39.3 (5.2)
	$\mathrm{Er}_{ng}(\widehat{\Theta})$	17.5 (5.3)	22.7 (7.0)	33.6 (10.5)	27.1 (4.6)	36.8 (6.9)	57.9 (11.1)	87.4 (11.1)
	$\mathrm{Er}(\widehat{\boldsymbol{\phi}})$	2.3 (3.0)	8.3 (7.2)	8.0 (8.2)	4.2 (4.3)	14.5 (8.4)	14.7 (8.4)	7.5 (11.7)
	Rank	2.0 (0.0)	1.9 (0.3)	1.9 (0.3)/2.0 (0.2)	6.6 (0.5)	6.9 (0.3)	4.0 (0.5)/5.5 (0.5)	–
	Ratio_*	1.1 (0.0)	1.1 (0.0)	1.2 (0.1)	0.9 (0.1)	0.9 (0.1)	1.2 (0.1)	1.0 (0.1)

Table 3
Simulation: results of Model 2. The layout is the same as that of Table 2.

M%		ORE.r	mRRR.r	gRRR.r	ORE.n	mRRR.n	gRRR.n	uGLM
0%	$\mathrm{Er}_g(\widehat{\Theta})$	2.9 (0.5)	2.9 (0.5)	3.0 (0.5)	4.5 (0.7)	4.5 (0.7)	5.2 (0.5)	9.9 (1.2)
	$\mathrm{Er}_{ng}(\widehat{\Theta})$	22.6 (1.6)	22.6 (1.6)	24.8 (2.6)	33.5 (2.6)	33.5 (2.6)	35.2 (2.9)	64.9 (9.2)
	$\mathrm{Er}(\widehat{\boldsymbol{\phi}})$	0.9 (1.8)	0.9 (1.8)	0.9 (1.7)	1.2 (2.4)	1.2 (2.4)	1.0 (1.9)	1.8 (3.9)
	Rank	2.0 (0.0)	2.0 (0.0)	2.0 (0.1)/1.9 (0.3)	9.7 (0.4)	9.7 (0.4)	5.5 (0.6)/8.6 (0.6)	–
	Ratio_*	1.0 (0.0)	1.0 (0.0)	1.1 (0.1)	1.2 (0.2)	1.2 (0.2)	1.2 (0.2)	0.8 (0.1)
10%	$\mathrm{Er}_g(\widehat{\Theta})$	2.7 (0.6)	2.8 (0.6)	3.0 (0.7)	4.2 (0.8)	4.6 (0.8)	5.5 (0.7)	10.5 (1.4)
	$\mathrm{Er}_{ng}(\widehat{\Theta})$	22.1 (1.6)	23.2 (1.7)	25.3 (2.2)	33.2 (2.9)	36.2 (3.1)	38.1 (3.4)	68.9 (9.9)
	$\mathrm{Er}(\widehat{\boldsymbol{\phi}})$	0.7 (1.6)	1.0 (2.1)	1.1 (2.1)	1.0 (2.4)	1.3 (2.3)	2.6 (3.8)	1.7 (3.9)
	Rank	2.0 (0.0)	2.0 (0.0)	2.0 (0.1)/2.0 (0.2)	9.8 (0.4)	9.8 (0.4)	5.5 (0.6)/8.8 (0.6)	–
	Ratio_*	1.0 (0.0)	1.0 (0.0)	1.1 (0.1)	1.2 (0.1)	1.2 (0.2)	1.3 (0.2)	0.8 (0.1)
20%	$\mathrm{Er}_g(\widehat{\Theta})$	2.8 (0.5)	3.0 (0.6)	3.2 (0.7)	4.4 (0.7)	5.2 (0.8)	6.0 (0.6)	11.1 (1.6)
	$\mathrm{Er}_{ng}(\widehat{\Theta})$	22.4 (1.8)	25.0 (2.3)	27.6 (2.7)	33.3 (2.7)	39.8 (3.4)	41.9 (3.8)	69.9 (10.6)
	$\mathrm{Er}(\widehat{\boldsymbol{\phi}})$	0.7 (1.5)	1.2 (1.8)	1.3 (1.8)	1.0 (2.5)	3.5 (4.5)	5.8 (6.6)	1.9 (4.4)
	Rank	2.0 (0.0)	2.0 (0.0)	2.0 (0.0)/2.0 (0.3)	9.7 (0.4)	9.9 (0.3)	5.9 (0.6)/9.0 (0.6)	–
	Ratio_*	1.0 (0.0)	1.0 (0.0)	1.1 (0.1)	1.2 (0.2)	1.3 (0.2)	1.3 (0.2)	0.8 (0.1)

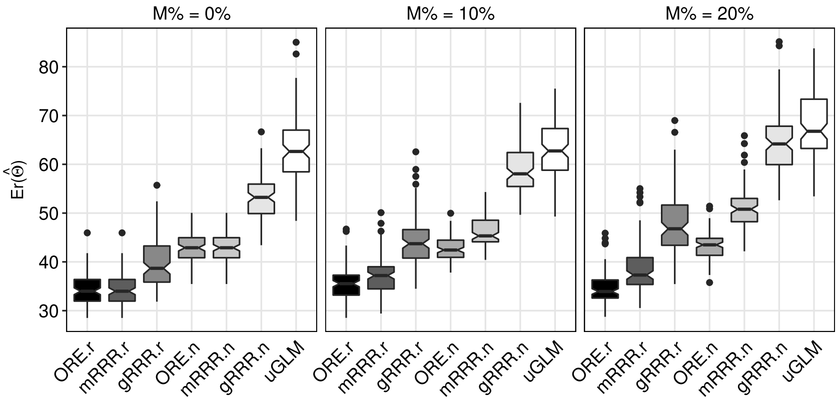


Fig. 1. Simulation: boxplots of the estimation error, $\mathrm{Er}(\widehat{\Theta})$, from the simulation results of Model 1.

outcomes; there are 41 binary outcomes which fall into several categories: 7 measures on fundamental daily activity, 13 on extended daily activity, 5 on social involvement, 8 on medical condition, 4 on cognitive ability, and 4 on sensation condition. In total, 20.2% of outcome values are missing.

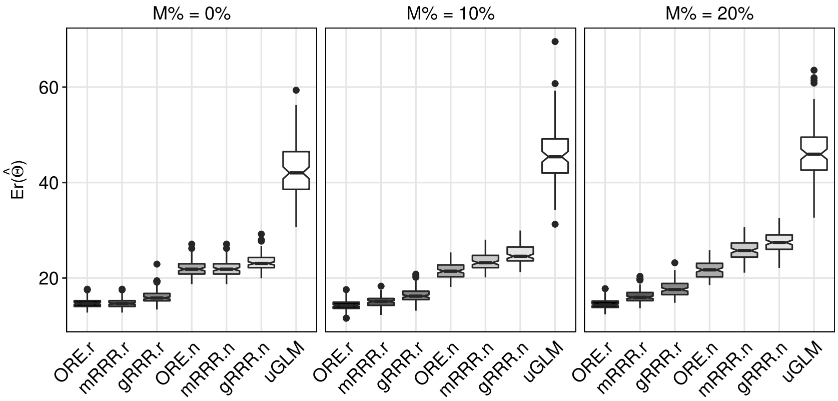


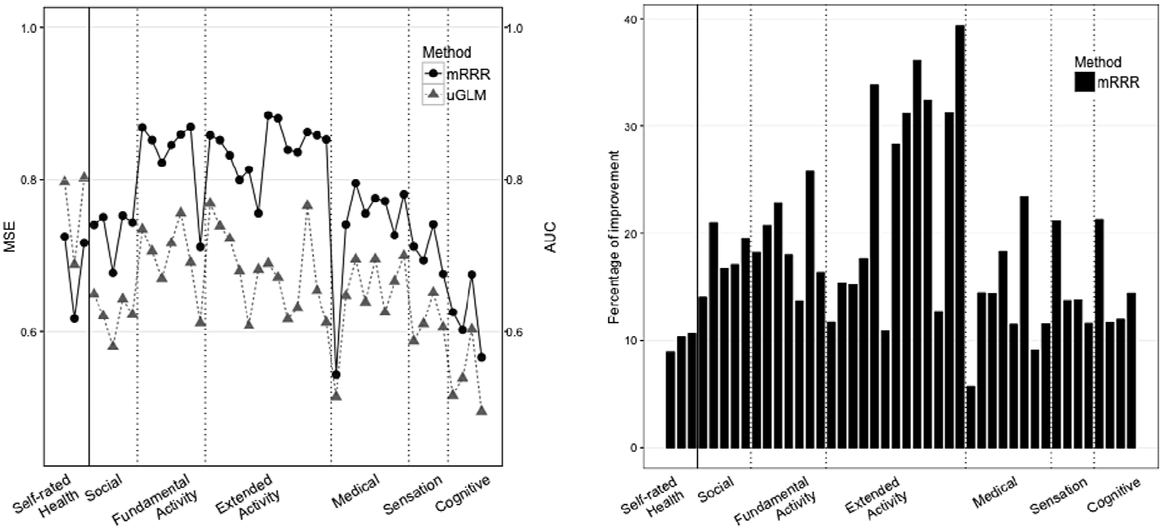
Fig. 2. Simulation: boxplots of the estimation error, Er($\widehat{\Theta}$), from the simulation results of model 2.

Potential predictors from 1997–1998 data include records of demographics, family structure, daily personal care, medical history, social activity, health opinion, behavior, nutrition, health insurance and income and assets, the majority of which are binary measurements. Among these variables, there are 13.7% missing values due to non-response and questionnaire filtering. For a few continuous predictors, the missing values are imputed with sample mean. For binary predictors, a better approach is to treat missing as a third category as it may also carry important information; as such, two dummy variables are created from each binary predictor with missing values (the third one is not necessary.) This results in $p = 294$ predictors.

There may be strong correlations among the large number of outcomes and the features. Therefore, it is plausible that the outcomes are dependent on the features only through a few latent factors, making dimension reduction and reduced-rank models applicable. We thus apply mRRR with rank penalization to conduct joint analysis of both the continuous and binary outcomes; as mRRR can deal with the missing values in the outcomes, neither data removal nor imputation is needed. The gender and age variables are used as control variables and their corresponding coefficients are not penalized. To demonstrate the efficacy of joint modeling, we mainly compare mRRR with the univariate approach uGLM. We use a random splitting procedure to evaluate the predictive performance. In each split, 75% of data are randomly selected for training and the rest 25% data for testing. The prediction of the continuous outcomes in each split is evaluated by mean squared errors (MSE), while the prediction of the binary outcomes is evaluated by Area Under Curve (AUC), based on testing data alone. The procedure is repeated 100 times.

We compute the average predictive measures from random splitting. The average MSE for Gaussian outcomes are 0.69 (0.06) and 0.76 (0.07), and the average AUC for binary outcomes are 0.77 (0.10) and 0.65 (0.11), for mRRR and uGLM, respectively (the standard deviations are reported in the parenthesis). The uGLM approach is outperformed by mRRR by a large margin, indicating the strength of low-rank estimation. Fig. 3 provides a more detailed performance comparison on the prediction of each individual outcome. It can be seen that the improvement by mRRR over uGLM is persistent across all the outcomes. The greatest improvement appears to be in the categories of fundamental activity and extended activity, where the percentage of improvement is over 20% for several outcomes. Indeed, these outcomes tend to be moderately or highly correlated, making joint estimation particularly beneficial. The mRRR also performs substantially better in predicting the three continuous responses related to self-rated health. This can be explained by the fact that two out of the three self-rated health measures are about memory and depression, which are very relevant to the variables in the category of cognition [23]. We have also tried the regularized version of uGLM using elastic net (uGLM-EN), whose average MSE for Gaussian outcomes and average AUC for binary outcomes are 0.70 (0.06) and 0.75 (0.11), respectively. As such, the performance of mRRR is only slightly better than that of uGLM-EN. This shows that shrinkage and sparse estimation could also be quite effective in this application, so a joint sparse and reduced-rank method may lead to even better performance (to be discussed in Section 7).

In our theoretical analysis in Section 3, we have assumed the independence of the error terms, which implies that the dependency of the responses can be fully captured by the low-rank structure of the natural parameter matrix. It is important to access the validity of this assumption in this application. Based on a referee's suggestion, we use the nonparametric test for independence proposed in Fan et al. [18] which extends the distance correlation test by Székely et al. [47] to the case of testing mutual independence between $q \geq 2$ random variables or vectors. The method is implemented in the R package `IndependenceTests`, but unfortunately, it cannot be directly applied due to the presence of missing values in our problem. We have thus adopted a naive multiple imputation approach. Specifically, for each Gaussian response the missing values are imputed by the observed mean value, and for each binary response the missing values are imputed by random Bernoulli draws with the probability of 1 being the observed mean value. The mRRR model is then fitted with the fully imputed data and the independence test is conducted with the obtained residuals from the q response variables. This procedure is repeated 100 times. We find that the p-values are stably around 0.08, with a standard deviation of 0.008. This result suggests that there is no apparent violation of the independent-error assumption in this particular application. We point out that it would be



(a) Prediction performance.

(b) Percentage of improvement over uGLM.

Fig. 3. LSOA data: comparison of predictive performance of mRRR and uGLM. Use 75% sample as training set. The left panel displays the MSE or AUC value for predicting each individual outcome. The right panel shows the percentage of improvement by mRRR over uGLM. In the left panel, the left axis shows MSE while the right axis shows AUC. In both panels, the solid vertical line separates the Gaussian and the binary outcomes, while the dashed vertical lines separate the binary outcomes to different categories.

interesting to further extend the mRRR model to capture the potential response correlation even after conditioning on the predictors; see Section 7 for more discussion.

We have also tried the gRRR approach, which fits the three Gaussian responses and the 41 binary responses separately. Using the aforedescribed random-splitting procedure with 75% samples for training, the gRRR approach yields almost identical results comparing to mRRR. That the benefit of joint modeling is not observed in this case is partly due to the fact that the number of Gaussian responses is quite small and the sample size is very large. We have then tried smaller sample sizes for training. From the random-splitting procedure with 25% data for training, the average MSE for Gaussian outcomes are 0.78 (0.06) and 0.80 (0.06), and the average AUC for binary outcomes are 0.75 (0.10) and 0.73 (0.10), for mRRR and gRRR, respectively; with 10% data for training, the average MSE for Gaussian outcomes are 0.86 (0.09) and 0.92 (0.09), and the average AUC for binary outcomes are 0.73 (0.11) and 0.70 (0.10), for mRRR and gRRR, respectively. Therefore, as the sample size becomes smaller, while the performance of both methods deteriorates, the gain of mRRR over gRRR becomes more revealing. The results suggest that integrative modeling can be quite effective especially when sample size is small or information from each individual response is limited.

7. Discussion

We will explore several extensions of mRRR that are of immediate interest in real applications. First, we could add a sparsity-inducing penalty on **C**, e.g., a row-wise group lasso penalty [52], to conduct simultaneous rank reduction and variable selection. Second, the model of the natural parameters in (2) can be extended to

$$\theta_{ik} = o_{ik} + \mathbf{x}_i^\top \mathbf{c}_k + \mathbf{z}_i^\top \boldsymbol{\beta}_k + s_{ik},$$

with $(i, k) \in \Omega$. Here, each s_{ik} is called a natural-shift parameter and $\mathbf{S} = (s_{ik})_{n\times q}$ is termed the natural-shift matrix, characterizing the additional effects in outcomes that cannot be explained by the linear function of **X** and **Z**. Certain low-dimensional assumptions on **S** are necessary to ensure identifiability. For example, when **S** is a unit-rank matrix, the model implies that all the y_{ik}s are related through another unsupervised latent feature, in addition to the supervised latent features from **X**. This setup then induces response correlation even after conditioning on **X**. Following She and Chen [43], **S** can also be assumed to be a sparse matrix for outlier detection, i.e., an entry s_{ik} is zero if the corresponding observation is a "normal observation", so that its natural parameter θ_{ik} is modeled in the usual way; otherwise, if an observation is an outlier, s_{ik} may be nonzero to capture its outlying effect. As such, **S** is assumed to be a sparse matrix to adjust for the outliers, so that the model estimation can be immune from potential data corruption. The model estimation can still be conducted via penalized log-likelihood, by adding additional regularization terms on **S**.

We have used an iterative singular value thresholding algorithm for handling mixed and incomplete outcomes. It is interesting to improve it by considering algorithmic acceleration techniques. Indeed, empirically we have tested that properly updating the scaling factor during iterations can substantially speed up computation. Besides rank and nuclear norm penalization, some other non-convex or hybrid penalization methods could be attractive [20,21,35]. In our theoretical analysis, we did not address the estimation problem of the dispersion parameters. This is known to be difficult in general in regularized estimation, and we will certainly explore this important issue in the future. Several authors studied the effective degrees of freedom of nuclear norm penalized estimation and reduced-rank estimation under Stein's unbiased risk estimation framework; see, e.g., [34,54]. It would be interesting to study this problem for the mRRR approach, which can then advance the development of information criteria for model evaluation and selection. Last but not the least, in our work the missing data are still regarded as happening at random, as we have not attempted to model its potential dependence with observed data. When explicit information regarding missing becomes available, it would be interesting and challenging to consider how to incorporate more general missing mechanisms in the high dimensional regression problem.

Acknowledgments

Chen's research was partially supported by the National Science Foundation grants DMS-1613295 and IIS-1718798. Wang's research was partially supported by the National Science Foundation grants IIS-1750326 and IIS-1716432. The authors are grateful to the Editor-in-Chief, the Associate Editor, and the two referees for their valuable comments and suggestions, which have led to significant improvement of the paper.

Appendix A. Proof of Theorem 1

Let $\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$ be the eigen-decomposition of $\mathcal{P}_{\mathbf{A}}$. Since $\mathcal{P}_{\mathbf{A}}$ is the projection matrix on the column space of $\mathbf{A}$, only the first $r(\mathbf{A})$ entries of $\boldsymbol{\Lambda}$ on the diagonal equal to 1, and all the remaining entries equal to 0. Then for any matrix $\mathbf{Q} \in \mathbb{R}^{n\times q}$, $\boldsymbol{\Lambda}\mathbf{U}^\top\mathbf{Q}$ can be written as an $r(\mathbf{A}) \times q$ matrix with non-zero entries on top of a $\{n - r(\mathbf{A})\} \times q$ matrix of zeros.

For incomplete data, denote the sampled sequence of entries by $(\omega_t)_{t=1}^s \in ([n] \times [q])^s$, where $\omega_t = (i_t, k_t) \in \Omega$ for all $t \in [s]$ and $\omega_1 \cup \cdots \cup \omega_s = \Omega$. Following [28], define a sequence of matrices $(\mathbf{E}_t)_{t=1}^s \in \mathbb{R}^{n\times q}$. Entries of $\mathbf{E}_t$ are all zeros except for the coefficient $\omega_t = (i_t, k_t)$ which is equal to 1, i.e., $(\mathbf{E}_t)_{i_t,k_t} = 1$. Then for $\epsilon_1, \ldots, \epsilon_s$, a Rademacher sequence independent from $(\omega_t, \mathbf{Y}_{(i_t,k_t)})_{t=1}^s$, we define $\boldsymbol{\Sigma}_R = (\epsilon_1\mathbf{E}_1 + \cdots + \epsilon_s\mathbf{E}_s)/s$. We first prove two lemmas.

Lemma A.1. *Suppose* Conditions 1 and 3 *hold. For the rank-penalized estimator in* (10)*, on the event*

$$\mathcal{A} = \left[(8e\mu\underline{\gamma})nq[\mathrm{E}\{d_1(\mathcal{P}_{\mathbf{A}}\boldsymbol{\Sigma}_R)\}]^2 + \frac{8\mu}{\underline{\gamma}}\, nqd_1^2\big(s^{-1}\mathcal{P}_{\mathbf{A}}\mathbf{E}\big) \le \lambda\right],$$

we have

$$\frac{\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F^2}{nq} \lesssim \max\left[\mu e K^2\sqrt{\frac{\ln(n+q)}{s}},\ \frac{\mu}{\underline{\gamma}}\,\{r(\mathbf{C}^*) + r(\mathbf{Z}\boldsymbol{\beta}^*)\}\lambda\right].$$

Proof. First, we denote by $\varepsilon(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}')$ the Bregman divergence [2,3], viz.

$$\varepsilon(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}') = \frac{1}{s}\left(\langle \mathbf{b}(\boldsymbol{\Theta}), \mathbf{1}_\Omega^{n\times q}\rangle_F - \langle \mathbf{b}(\boldsymbol{\Theta}'), \mathbf{1}_\Omega^{n\times q}\rangle_F - \langle P_\Omega\{\boldsymbol{\mu}(\boldsymbol{\Theta}')\}, \boldsymbol{\Theta} - \boldsymbol{\Theta}'\rangle_F\right).$$

By Taylor expansion, it follows that

$$\frac{\underline{\gamma}}{2s}\,\|P_\Omega(\boldsymbol{\Theta} - \boldsymbol{\Theta}')\|_F^2 \le \varepsilon(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}') \le \frac{\overline{\gamma}}{2s}\,\|P_\Omega(\boldsymbol{\Theta} - \boldsymbol{\Theta}')\|_F^2.$$

In the special case of Gaussian, for example, $\varepsilon(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}') = \|P_\Omega(\boldsymbol{\Theta} - \boldsymbol{\Theta}')\|_F^2/(2s)$. This connects Bregman divergence to the error metric we are interested in.

By the definition of $(\widehat{\mathbf{C}}, \widehat{\boldsymbol{\beta}})$ in (10),

$$-\frac{1}{s}\langle \widetilde{\mathbf{Y}}, \widehat{\boldsymbol{\Theta}}\rangle_F + \frac{1}{s}\langle \mathbf{b}(\widehat{\boldsymbol{\Theta}}), \mathbf{1}_\Omega^{n\times q}\rangle_F + \lambda r(\widehat{\mathbf{C}}) \le -\frac{1}{s}\langle \widetilde{\mathbf{Y}}, \boldsymbol{\Theta}\rangle_F + \frac{1}{s}\langle \mathbf{b}(\boldsymbol{\Theta}), \mathbf{1}_\Omega^{n\times q}\rangle_F + \lambda r(\mathbf{C}),$$

for all $p \times q$ matrices $\mathbf{C}$ with rank $\widehat{k}$ and $(p_z + 1) \times q$ matrices $\boldsymbol{\beta}$. Then with the definition $\mathbf{E} = \widetilde{\mathbf{Y}} - P_\Omega\{\boldsymbol{\mu}(\boldsymbol{\Theta}^*)\}$, we have

$$-\frac{1}{s}\langle P_\Omega\{\boldsymbol{\mu}(\boldsymbol{\Theta}^*)\} + \mathbf{E}, \widehat{\boldsymbol{\Theta}}\rangle_F + \frac{1}{s}\,\langle \mathbf{b}(\widehat{\boldsymbol{\Theta}}), \mathbf{1}_\Omega^{n\times q}\rangle_F + \lambda r(\widehat{\mathbf{C}}) \le -\frac{1}{s}\,\langle P_\Omega\{\boldsymbol{\mu}(\boldsymbol{\Theta}^*)\} + \mathbf{E}, \boldsymbol{\Theta}\rangle_F + \frac{1}{s}\,\langle \mathbf{b}(\boldsymbol{\Theta}), \mathbf{1}_\Omega^{n\times q}\rangle_F + \lambda r(\mathbf{C}).$$

By adding $s^{-1}\langle P_\Omega\{\boldsymbol{\mu}(\boldsymbol{\Theta}^*)\}, \boldsymbol{\Theta}^*\rangle_F - s^{-1}\langle \mathbf{b}(\boldsymbol{\Theta}^*), \mathbf{1}_\Omega^{n\times q}\rangle_F$ to both sides, we have

$$\begin{aligned}\varepsilon(\widehat{\boldsymbol{\Theta}} \mid \boldsymbol{\Theta}^*) &\le \varepsilon(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^*) + 2\lambda r(\mathbf{C}) + 2\langle s^{-1}\mathbf{E}, \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\rangle_F - \lambda r(\widehat{\mathbf{C}}) - \lambda r(\mathbf{C}) \\ &= \varepsilon(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^*) + 2\lambda r(\mathbf{C}) + 2\langle s^{-1}\,\mathcal{P}_{\mathbf{A}}\mathbf{E}, \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\rangle_F - \lambda r(\widehat{\mathbf{C}}) - \lambda r(\mathbf{C}).\end{aligned}$$

Let $\mathbf{C} = \mathbf{C}^*$, $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, we have

$$\varepsilon(\widehat{\boldsymbol{\Theta}} \mid \boldsymbol{\Theta}^*) \le 2\lambda r(\mathbf{C}^*) + 2\langle s^{-1}\mathcal{P}_{\mathbf{A}}\mathbf{E}, \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\rangle_F - \lambda r(\widehat{\mathbf{C}}) - \lambda r(\mathbf{C}^*). \tag{13}$$

The inner product term can be bounded as follows,

$$\begin{aligned}\langle s^{-1}\mathcal{P}_{\mathbf{A}}\mathbf{E}, \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\rangle_F &\le d_1\big(s^{-1}\mathcal{P}_{\mathbf{A}}\mathbf{E}\big)\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_* \\ &= d_1\big(s^{-1}\mathcal{P}_{\mathbf{A}}\mathbf{E}\big)\|\mathbf{X}\widehat{\mathbf{C}} + \mathbf{Z}\widehat{\boldsymbol{\beta}} - (\mathbf{X}\mathbf{C}^* + \mathbf{Z}\boldsymbol{\beta}^*)\|_* \\ &\le d_1\big(s^{-1}\mathcal{P}_{\mathbf{A}}\mathbf{E}\big)\{r(\mathbf{X}\widehat{\mathbf{C}}) + r(\mathbf{X}\mathbf{C}^*) + r(\mathbf{Z}\widehat{\boldsymbol{\beta}}) + r(\mathbf{Z}\boldsymbol{\beta}^*)\}^{1/2}\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F \\ &\le d_1\big(s^{-1}\mathcal{P}_{\mathbf{A}}\mathbf{E}\big)\{r(\widehat{\mathbf{C}}) + r(\mathbf{C}^*) + 2r(\mathbf{Z}\boldsymbol{\beta}^*)\}^{1/2}\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F, \end{aligned} \tag{14}$$

where we assume $\boldsymbol{\beta}^*$ has full rank.

By (13) and (14), and using the inequality $2xy \le x^2/a + ay^2$, for any $a > 0$, we have

$$\varepsilon(\widehat{\boldsymbol{\Theta}} \mid \boldsymbol{\Theta}^*) \le \frac{1}{a}\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F^2 + \{ad_1^2\big(s^{-1}\mathcal{P}_{\mathbf{A}}\mathbf{E}\big) - \lambda\}\{r(\widehat{\mathbf{C}}) + r(\mathbf{C}^*)\} + 2\lambda r(\mathbf{C}^*) + 2r(\mathbf{Z}\boldsymbol{\beta}^*)ad_1^2\big(s^{-1}\mathcal{P}_{\mathbf{A}}\mathbf{E}\big). \tag{15}$$

Denote

$$\Delta_\Omega^2(\widehat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}^*) = \frac{1}{s}\|P_\Omega(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*)\|_F^2 \le \frac{2}{\underline{\gamma}}\,\varepsilon(\widehat{\boldsymbol{\Theta}} \mid \boldsymbol{\Theta}^*). \tag{16}$$

Consider the following two cases according to a threshold value $8eK^2\sqrt{\ln\{r(\mathbf{A}) + q\}/s}$.

Case 1: $\sum_{ik}\pi_{ik}(\widehat{\theta}_{ik} - \theta_{ik}^*)^2 \le 8eK^2\sqrt{\ln\{r(\mathbf{A}) + q\}/s}$. Using Condition 3, we have

$$\frac{\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F^2}{nq} \le \sum_{ik}\pi_{ik}(\widehat{\theta}_{ik} - \theta_{ik}^*)^2 \le 8\mu eK^2\sqrt{\ln\{r(\mathbf{A}) + q\}/s}. \tag{17}$$

We thus obtain the first term in the final bound.

Case 2: $\sum_{ik}\pi_{ik}(\widehat{\theta}_{ik} - \theta_{ik}^*)^2 > 8eK^2\sqrt{\ln\{r(\mathbf{A}) + q\}/s}$. We have

$$\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_* \le \{r(\widehat{\mathbf{C}}) + r(\mathbf{C}^*) + 2r(\mathbf{Z}\boldsymbol{\beta}^*)\}^{1/2}\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F \le \sqrt{\{r(\widehat{\mathbf{C}}) + r(\mathbf{C}^*) + 2r(\mathbf{Z}\boldsymbol{\beta}^*)\}(\mu nq)\mathrm{E}\{\Delta_\Omega^2(\widehat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}^*)\}},$$

where $\mathrm{E}\{\Delta_\Omega^2(\widehat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}^*)\} = \sum_{ik}\pi_{ik}(\widehat{\theta}_{ik} - \theta_{ik}^*)^2$. Also, using Lemma 19 of [28], with probability at least $1 - \{r(\mathbf{A}) + q - 1\}^{-1} \ge 1 - 2\{r(\mathbf{A}) + q\}^{-1}$,

$$\Delta_\Omega^2(\widehat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}^*) \ge \frac{1}{2}\mathrm{E}\{\Delta_\Omega^2(\widehat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}^*)\} - 16e[\mathrm{E}\{d_1(\mathcal{P}_{\mathbf{A}}\boldsymbol{\Sigma}_R)\}]^2(\mu nq)\{r(\widehat{\mathbf{C}}) + r(\mathbf{C}^*) + 2r(\mathbf{Z}\boldsymbol{\beta}^*)\}. \tag{18}$$

Now, combining the results in (15), (16) and (18) we have

$$\begin{aligned}&\frac{\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F^2}{2\mu nq} - 16e[\mathrm{E}\{d_1(\mathcal{P}_{\mathbf{A}}\boldsymbol{\Sigma}_R)\}]^2(\mu nq)\{r(\widehat{\mathbf{C}}) + r(\mathbf{C}^*) + 2r(\mathbf{Z}\boldsymbol{\beta}^*)\} \\ &\le \frac{2}{\underline{\gamma}}\left[\frac{1}{a}\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F^2 + \Big\{ad_1^2\big(s^{-1}\mathcal{P}_{\mathbf{A}}\mathbf{E}\big) - \lambda\Big\}\{r(\widehat{\mathbf{C}}) + r(\mathbf{C}^*)\} + 2\lambda r(\mathbf{C}^*) + 2r(\mathbf{Z}\boldsymbol{\beta}^*)ad_1^2\big(s^{-1}\mathcal{P}_{\mathbf{A}}\mathbf{E}\big)\right].\end{aligned}$$

By choosing $a = (8\mu nq)/\underline{\gamma}$, we have

$$\begin{aligned}\frac{\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F^2}{nq}\left(\frac{1}{2\mu} - \frac{1}{4\mu}\right) &\le \left[(16e\mu)nq[\mathrm{E}\{d_1(\mathcal{P}_{\mathbf{A}}\boldsymbol{\Sigma}_R)\}]^2 + \frac{16\mu nq}{\underline{\gamma}^2}d_1^2\big(s^{-1}\mathcal{P}_{\mathbf{A}}\mathbf{E}\big) - \frac{2}{\underline{\gamma}}\lambda\right]\{r(\widehat{\mathbf{C}}) + r(\mathbf{C}^*)\} \\ &\quad + \left[(16e\mu)nq[\mathrm{E}\{d_1(\mathcal{P}_{\mathbf{A}}\boldsymbol{\Sigma}_R)\}]^2 + \frac{16\mu nq}{\underline{\gamma}^2}d_1^2\big(s^{-1}\mathcal{P}_{\mathbf{A}}\mathbf{E}\big)\right]2r(\mathbf{Z}\boldsymbol{\beta}^*) + \frac{4}{\underline{\gamma}}\lambda r(\mathbf{C}^*).\end{aligned}$$

It follows that on the event $\mathcal{A}$,

$$\frac{\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F^2}{nq} \lesssim \frac{\mu\lambda}{\underline{\gamma}}\{r(\mathbf{C}^*) + r(\mathbf{Z}\boldsymbol{\beta}^*)\}. \tag{19}$$

The proof is completed by combining the bounds in (17) and (19). □

Lemma A.2. *Suppose Conditions 1–4 hold. For a constant $\alpha > 0$, we have*

$$\mathrm{E}\{d_1(\mathcal{P}_{\mathbf{A}}\boldsymbol{\Sigma}_R)\} \le \alpha\sqrt{\frac{2e\nu\ln\{r(\mathbf{A}) + q\}}{(n \wedge q)s}}.$$

For a constant $c_E > 0$ which depends on σ_E, we have, with probability of at least $1 - \{r(\mathbf{A}) + q\}^{-1}$,

$$d_1^2\big(s^{-1}\mathcal{P}_{\mathbf{A}}\mathbf{E}\big) \le c_E^2\overline{\gamma}\,\frac{2\ln\{r(\mathbf{A}) + q\}\nu}{(n \wedge q)s}.$$

Proof. (a) Bounding $\mathrm{E}\{d_1(\mathcal{P}_{\mathbf{A}}\boldsymbol{\Sigma}_R)\}$: First, one has $d_1(\mathcal{P}_{\mathbf{A}}\boldsymbol{\Sigma}_R) \leq d_1(\boldsymbol{\Lambda}\mathbf{U}^\top\boldsymbol{\Sigma}_R)$. One can write $\boldsymbol{\Sigma}'_R = (\mathbf{W}_1 + \cdots + \mathbf{W}_s)/s$, with $\mathbf{W}_t$ denotes the $r(\mathbf{A}) \times q$ non-zero matrix of $\epsilon_t\boldsymbol{\Lambda}\mathbf{U}^\top\mathbf{E}_t$ and satisfies $\mathrm{E}(\mathbf{W}_t) = \mathbf{0}$. Then we have $\mathrm{E}\{d_1(\boldsymbol{\Lambda}\mathbf{U}^\top\boldsymbol{\Sigma}_R)\} = \mathrm{E}\{d_1(\boldsymbol{\Sigma}'_R)\}$. Denoting $R_i = \pi_{i1} + \cdots + \pi_{iq}$ for each $i \in \{1, \ldots, n\}$, one obtains

$$d_1\left\{\mathrm{E}\left(\frac{1}{s}\sum_{t=1}^{s}\mathbf{W}_t\mathbf{W}_t^\top\right)\right\} \leq d_1\left\{\mathrm{E}\left(\frac{1}{s}\sum_{t=1}^{s}\epsilon_t^2\mathbf{E}_t\mathbf{E}_t^\top\right)\right\} \leq d_1\{\mathrm{diag}(R_1, \ldots, R_n)\} \leq \frac{\nu}{n \wedge q},$$

where Condition 3 was used for the last inequality. Using a similar argument one also gets $d_1\{\mathrm{E}(s^{-1}\sum_{t=1}^{s}\mathbf{W}_t^\top\mathbf{W}_t)\} \leq \nu/n \wedge q$. Hence applying Lemma 20 of Lafond [28] with $m_1 = r(\mathbf{A})$, $m_2 = q$, $n = s$, $Z_i = \mathbf{W}_t$, $U = 1$, $c^* = \alpha = 1 + \sqrt{3}$, $d = r(\mathbf{A}) + q$ and $\sigma_Z^2 = \nu/n \wedge q$, for $s > \ln\{r(\mathbf{A}) + q\}(n \wedge q)/(9\nu)$ yields

$$\mathrm{E}\{d_1(\boldsymbol{\Sigma}'_R)\} \leq \alpha\sqrt{\frac{2e\nu\ln\{r(\mathbf{A}) + q\}}{(n \wedge q)s}}.$$

(b) Bounding $d_1(s^{-1}\mathcal{P}_{\mathbf{A}}\mathbf{E})$: First, one has $d_1(s^{-1}\mathcal{P}_{\mathbf{A}}\mathbf{E}) \leq d_1(s^{-1}\boldsymbol{\Lambda}\mathbf{U}^\top\mathbf{E})$. Let us define $\mathbf{W}'_t$ as the $r(\mathbf{A}) \times q$ non-zero matrix of $\{y_{i_t,k_t} - \mu(\theta^*_{i_t,k_t})\}\boldsymbol{\Lambda}\mathbf{U}^\top\mathbf{E}_t$, which satisfies $\mathrm{E}(\mathbf{W}'_t) = \mathbf{0}$ (as any score function) and

$$\sigma_{\mathbf{W}'}^2 = \max\left[\frac{1}{s}\,d_1\left[\mathrm{E}\left\{\sum_{t=1}^{s}(\mathbf{W}'_t)^\top\mathbf{W}'_t\right\}\right], \frac{1}{s}d_1\left[\mathrm{E}\left\{\sum_{t=1}^{s}\mathbf{W}'_t(\mathbf{W}'_t)^\top\right\}\right]\right].$$

Then we have $d_1(s^{-1}\boldsymbol{\Lambda}\mathbf{U}^\top\mathbf{E}) = d_1\{(\mathbf{W}'_1 + \cdots + \mathbf{W}'_s)/s\}$. Using Conditions 1 and 2, a similar analysis yields $\sigma_{\mathbf{W}'}^2 \leq \overline{\gamma}\nu/n \wedge q$.

Combining $\max_{i,k}(\sum_i\pi_{ik}, \sum_k\pi_{ik}) \geq 1/n \wedge q$ and $\mathrm{E}[\{y_{i_t,k_t} - \mu(\theta^*_{i_t,k_t})\}^2] = b''(\theta^*_{i_t,k_t}) \geq \underline{\gamma}$, we also have $\sigma_{\mathbf{W}'}^2 \geq \underline{\gamma}/n \wedge q$. Since

$$d_1(\mathbf{W}'_t) = d_1[\{y_{i_t,k_t} - \mu(\theta^*_{i_t,k_t})\}\boldsymbol{\Lambda}\mathbf{U}^\top\mathbf{E}_t] \leq d_1[\{y_{i_t,k_t} - \mu(\theta^*_{i_t,k_t})\}\mathbf{E}_t],$$

applying Proposition 21 of Lafond [28] for $m_1 = r(\mathbf{A})$, $m_2 = q$, $t = \ln\{r(\mathbf{A}) + q\}$, $Z_i = \mathbf{W}'_t$, $U = \sigma_E$ and $\sigma_Z = \sigma_{\mathbf{W}'}$ gives with probability at least $1 - \{r(\mathbf{A}) + q\}^{-1}$

$$d_1\left(\frac{1}{s}\sum_{t=1}^{s}\mathbf{W}'_t\right) \leq c_E\max\left[\sqrt{\overline{\gamma}\,\frac{2\ln\{r(\mathbf{A}) + q\}\nu}{(n \wedge q)s}}, \sigma_E\ln\left(\frac{\sigma_E\sqrt{n \wedge q}}{\sqrt{\underline{\gamma}}}\right)\frac{2\ln\{r(\mathbf{A}) + q\}}{s}\right],$$

with c_E which depends only on σ_E. By assumption on s, the left term dominates. □

Now we finish the proof of Theorem 1. We have

$$\begin{aligned}
&(8e\mu\underline{\gamma})nq\,[\mathrm{E}\{d_1(\mathcal{P}_{\mathbf{A}}\boldsymbol{\Sigma}_R)\}]^2 + \frac{8\mu}{\underline{\gamma}}\,nqd_1^2\big(s^{-1}\mathcal{P}_{\mathbf{A}}\mathbf{E}\big)\\
&\leq 16\mu\nu\alpha^2e^2\underline{\gamma}nq\frac{\ln\{r(\mathbf{A}) + q\}}{(n \wedge q)s} + \frac{16\mu\nu}{\underline{\gamma}}c_E^2\overline{\gamma}nq\frac{\ln\{r(\mathbf{A}) + q\}}{(n \wedge q)s}\\
&= 16\mu\nu\left(\alpha^2e^2\underline{\gamma} + c_E^2\,\frac{\overline{\gamma}}{\underline{\gamma}}\right)\frac{\ln\{r(\mathbf{A}) + q\}}{s}(n \vee q)
\end{aligned}$$

with probability at least $1-\{r(\mathbf{A})+q\}^{-1}$. The results immediately follow by combining the results in Lemmas A.1 and A.2. □

Appendix B. Details on computation

Derivation of Algorithm 1

We acknowledge that the derivation of the proposed algorithm follows similar architecture as in She [42]. We generalize the method to handle incomplete data and mixed outcomes. For updating $\mathbf{C}$, consider a surrogate function of $F(\mathbf{C}, \boldsymbol{\beta}, \boldsymbol{\phi}_u)$ in (3), viz.

$$\begin{aligned}
G(\mathbf{A}; \mathbf{C}, \boldsymbol{\beta}, \boldsymbol{\phi}_u) = &- \sum_{(i,k)\in\Omega}\ell_k(\mathbf{a}_k, \boldsymbol{\beta}_k, \phi_k; \mathbf{x}_i, \mathbf{z}_i, y_{ik}) + \sum_{h=1}^{p\wedge q}\rho\{d_h(\mathbf{A}); \lambda\} + \frac{1}{2}\,\|\mathbf{A} - \mathbf{C}\|_F^2\\
&- \sum_{(i,k)\in\Omega}\{b_k(\mathbf{x}_i^\top\mathbf{a}_k + \mathbf{z}_i^\top\boldsymbol{\beta}_k) - b_k(\mathbf{x}_i^\top\mathbf{c}_k + \mathbf{z}_i^\top\boldsymbol{\beta}_k)\}/a_k(\phi_k)\\
&+ \sum_{(i,k)\in\Omega}g_k^{-1}(\mathbf{x}_i^\top\mathbf{c}_k + \mathbf{z}_i^\top\boldsymbol{\beta}_k)(\mathbf{x}_i^\top\mathbf{a}_k - \mathbf{x}_i^\top\mathbf{c}_k)/a_k(\phi_k),
\end{aligned} \tag{20}$$

where ℓ_k is as defined in (4). It is easy to see that $G(\mathbf{C}; \mathbf{C}, \boldsymbol{\beta}, \boldsymbol{\phi}_u) = F(\mathbf{C}, \boldsymbol{\beta}, \boldsymbol{\phi}_u)$. After some algebra, G can be simplified as

$$G(\mathbf{A}; \mathbf{C}, \boldsymbol{\beta}, \boldsymbol{\phi}_u) = \frac{1}{2}\,\|\mathbf{A} - \mathbf{C} - \mathbf{X}^\top P_\Omega\{\mathbf{Y} - \boldsymbol{\mu}(\mathbf{C}, \boldsymbol{\beta})\}\boldsymbol{\Phi}^{-1}\|_F^2 + \sum_{h=1}^{p\wedge q}\rho\{d_h(\mathbf{A}); \lambda\} + \mathrm{const}, \tag{21}$$

where "const" represents any remainder constant term that does not depend on $\mathbf{A}$. A core setup in our algorithm is to minimize $G(\mathbf{A};\mathbf{C},\boldsymbol{\beta},\boldsymbol{\phi}_u)$ with respect to $\mathbf{A}$. At the tth iteration, when $\mathbf{C}=\mathbf{C}^{(t)}$, $\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}$ and $\boldsymbol{\phi}=\boldsymbol{\phi}^{(t)}$, the minimizer of $G(\mathbf{A};\mathbf{C}^{(t)},\boldsymbol{\beta}^{(t)},\boldsymbol{\phi}_u^{(t)})$ is

$$\widehat{\mathbf{A}}=\mathbf{C}^{(t+1)}=\mathbb{T}^d[\mathbf{C}^{(t)}+\mathbf{X}^\top P_\Omega\{\mathbf{Y}-\boldsymbol{\mu}(\mathbf{C}^{(t)},\boldsymbol{\beta}^{(t)})\}\boldsymbol{\Phi}^{(t)-1};\lambda],$$

following the results in (11) and (12). In Theorem 2, we show that under mild conditions, this update ensures the descending of the objective.

When $\mathbf{C}$ is held fixed, solving (3) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\phi}_u$ boils down to a set of univariate GLM problems. When non-Gaussian outcomes present, in general their corresponding GLM problems need iterative algorithms to solve, which could be very time consuming. Here, alternatively, we construct another surrogate function to get an one-step update of $\boldsymbol{\beta}$, similar to the previous updating of $\mathbf{C}$. Define

$$\begin{aligned}H(\boldsymbol{\alpha};\mathbf{C},\boldsymbol{\beta},\boldsymbol{\phi}_u)=&-\sum_{(i,k)\in\Omega}\ell_k(\mathbf{c}_k,\boldsymbol{\alpha}_k,\phi_k;\mathbf{x}_i,\mathbf{z}_i,y_{ik})+\sum_{h=1}^{p\wedge q}\rho\{d_h(\mathbf{C});\lambda_1\}+\frac{1}{2}\|\boldsymbol{\alpha}-\boldsymbol{\beta}\|_F^2\\&-\sum_{(i,k)\in\Omega}\{b_k(\mathbf{x}_i^\top\mathbf{c}_k+\mathbf{z}_i^\top\boldsymbol{\alpha}_k)-b_k(\mathbf{x}_i^\top\mathbf{c}_k+\mathbf{z}_i\boldsymbol{\beta}_k)\}/a_k(\phi_k)\\&+\sum_{(i,k)\in\Omega}g_k^{-1}(\mathbf{x}_i^\top\mathbf{c}_k+\mathbf{z}_i^\top\boldsymbol{\beta}_k)(\mathbf{z}_i^\top\boldsymbol{\alpha}_k-\mathbf{z}_i^\top\boldsymbol{\beta}_k)/a_k(\phi_k).\end{aligned}$$

Then minimizing H with respect to $\boldsymbol{\alpha}$ is the same as minimizing $\|\boldsymbol{\alpha}-\boldsymbol{\beta}-\mathbf{Z}^\top P_\Omega\{\mathbf{Y}-\boldsymbol{\mu}(\mathbf{C},\boldsymbol{\beta})\}\boldsymbol{\Phi}^{-1}\|_F^2$, which is a least squares problem. When $\mathbf{C}=\mathbf{C}^{(t+1)}$, $\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}$, and $\boldsymbol{\phi}=\boldsymbol{\phi}^{(t)}$, the minimizer is $\widehat{\boldsymbol{\alpha}}=\boldsymbol{\beta}^{(t+1)}=\boldsymbol{\beta}^{(t)}+\mathbf{Z}^\top P_\Omega\{\mathbf{Y}-\boldsymbol{\mu}(\mathbf{C}^{(t+1)},\boldsymbol{\beta}^{(t)})\}\boldsymbol{\Phi}^{(t)-1}$. Once $\mathbf{C}$ and $\boldsymbol{\beta}$ are updated, we can then update $\boldsymbol{\phi}_u$ by maximizing the log-likelihood function.

Proof of Theorem 2. Consider first the surrogate function G defined in (20). As shown in Proposition 2.2 of She [42], for any $\boldsymbol{\Delta}_1\in\mathbb{R}^{p\times q}$,

$$G(\mathbf{C}^{(t+1)}+\boldsymbol{\Delta}_1;\mathbf{C}^{(t)},\boldsymbol{\beta}^{(t)},\boldsymbol{\phi}_u^{(t)})-G(\mathbf{C}^{(t+1)};\mathbf{C}^{(t)},\boldsymbol{\beta}^{(t)},\boldsymbol{\phi}_u^{(t)})\geq\frac{\eta_1}{2}\|\boldsymbol{\Delta}_1\|_F^2,$$

where $\eta_1=\max(0,1-L_1)$, and $L_1\in[0,1]$ is a constant such that for the thresholding rule $\mathbb{T}$ corresponding to $\underline{\rho}$, $d\mathbb{T}^{-1}(u;\lambda)/du$ is bounded below by $1-L_1$. Using Taylor expansion,

$$\begin{aligned}&\sum_{(i,k)\in\Omega}\frac{1}{a_k(\phi_k^{(t)})}\{b_k(\mathbf{x}_i^\top\mathbf{c}_k^{(t+1)}+\mathbf{z}_i^\top\boldsymbol{\beta}_k^{(t)})-b_k(\mathbf{x}_i^\top\mathbf{c}_k^{(t)}+\mathbf{z}_i^\top\boldsymbol{\beta}_k^{(t)})\}\\&=\sum_{(i,k)\in\Omega}\frac{1}{a_k(\phi_k^{(t)})}\{g_k^{-1}(\mathbf{x}_i^\top\mathbf{c}_k^{(t)}+\mathbf{z}_i^\top\boldsymbol{\beta}_k^{(t)})(\mathbf{x}_i^\top\mathbf{c}_k^{(t+1)}-\mathbf{x}_i^\top\mathbf{c}_k^{(t)})\\&\quad+\frac{1}{2}b_k''(\mathbf{x}_i^\top\boldsymbol{\xi}_k^{(t+1)}+\mathbf{z}_i^\top\boldsymbol{\beta}_k^{(t)})(\mathbf{x}_i^\top\mathbf{c}_k^{(t+1)}-\mathbf{x}_i^\top\mathbf{c}_k^{(t)})^2\},\end{aligned}$$

where $\boldsymbol{\xi}_k^{(t+1)}\in\{a\mathbf{c}_k^{(t)}+(1-a)\mathbf{c}_k^{(t+1)};0<a<1\}$. It follows that

$$\begin{aligned}G(\mathbf{C}^{(t+1)};\mathbf{C}^{(t)},\boldsymbol{\beta}^{(t)},\boldsymbol{\phi}_u^{(t)})&=F(\mathbf{C}^{(t+1)},\boldsymbol{\beta}^{(t)},\boldsymbol{\phi}_u^{(t)})\\&\quad-\frac{1}{2}\sum_{(i,k)\in\Omega}\frac{1}{a_k(\phi_k^{(t)})}b_k''(\mathbf{x}_i^\top\boldsymbol{\xi}_k^{(t+1)}+\mathbf{z}_i^\top\boldsymbol{\beta}_k^{(t)})(\mathbf{x}_i^\top\mathbf{c}_k^{(t+1)}-\mathbf{x}_i^\top\mathbf{c}_k^{(t)})^2+\frac{1}{2}\|\mathbf{C}^{(t+1)}-\mathbf{C}^{(t)}\|_F^2\\&\leq G(\mathbf{C}^{(t)};\mathbf{C}^{(t)},\boldsymbol{\beta}^{(t)},\boldsymbol{\phi}_u^{(t)})-\frac{\eta_1}{2}\|\mathbf{C}^{(t+1)}-\mathbf{C}^{(t)}\|_F^2\\&=F(\mathbf{C}^{(t)},\boldsymbol{\beta}^{(t)},\boldsymbol{\phi}_u^{(t)})-\frac{\eta_1}{2}\|\mathbf{C}^{(t+1)}-\mathbf{C}^{(t)}\|_F^2.\end{aligned}$$

Therefore,

$$\begin{aligned}&F(\mathbf{C}^{(t)},\boldsymbol{\beta}^{(t)},\boldsymbol{\phi}_u^{(t)})-F(\mathbf{C}^{(t+1)},\boldsymbol{\beta}^{(t)},\boldsymbol{\phi}^{(t)})\\&\geq\frac{1+\eta_1}{2}\|\mathbf{C}^{(t+1)}-\mathbf{C}^{(t)}\|_F^2-\frac{1}{2}\sum_{(i,k)\in\Omega}\frac{1}{a_k(\phi_k^{(t)})}b_k''(\mathbf{x}_i^\top\boldsymbol{\xi}_k^{(t+1)}+\mathbf{z}_i^\top\boldsymbol{\beta}_k^{(t)})(\mathbf{x}_i^\top\mathbf{c}_k^{(t+1)}-\mathbf{x}_i^\top\mathbf{c}_k^{(t)})^2.\end{aligned}$$

The second term on the right-hand side is bounded from above by

$$\frac{1}{2}\sum_{k=1}^{q}(\mathbf{c}_k^{(t+1)}-\mathbf{c}_k^{(t)})^\top\mathbf{I}(\boldsymbol{\xi}_k^{(t+1)},\boldsymbol{\beta}_k^{(t)},\phi_k^{(t)};\mathbf{X})(\mathbf{c}_k^{(t+1)}-\mathbf{c}_k^{(t)}),$$

where we have defined

$$\mathbf{W}(\mathbf{c}_k, \boldsymbol{\beta}_k, \phi_k) = \operatorname{diag}\left\{ \frac{1}{a_k(\phi_k)} |b_k''(\mathbf{x}_1^\top \mathbf{c}_k + \mathbf{z}_1^\top \boldsymbol{\beta}_k)|, \ldots, \frac{1}{a_k(\phi_k)} |b_k''(\mathbf{x}_n^\top \mathbf{c}_k + \mathbf{z}_n^\top \boldsymbol{\beta}_k)| \right\},$$

and

$$\mathbf{I}(\mathbf{c}_k, \boldsymbol{\beta}_k, \phi_k; \mathbf{X}) = \sum_{i=1}^{n} \frac{1}{a_k(\phi_k)} |b_k''(\mathbf{x}_i^\top \mathbf{c}_k + \mathbf{z}_i^\top \boldsymbol{\beta}_k)| \mathbf{x}_i \mathbf{x}_i^\top = \mathbf{X}^\top \mathbf{W}(\mathbf{c}_k, \boldsymbol{\beta}_k, \phi_k) \mathbf{X}.$$

Let $\mathcal{A}_k = \{(a\mathbf{c}_k^{(t)} + (1-a)\mathbf{c}_k^{(t+1)}, \boldsymbol{\beta}_k^{(t)}, \phi_k^{(t)}) : a \in (0,1), t \in \{1,2,\ldots\}\}$, and

$$\gamma_1 = \max_{k \in \{1,\ldots,q\}} \sup_{(\boldsymbol{\xi}_k, \boldsymbol{\zeta}_k, \delta_k) \in \mathcal{A}_k} \|\mathbf{I}(\boldsymbol{\xi}_k, \boldsymbol{\zeta}_k, \delta_k)\|_2.$$

Then

$$F(\mathbf{C}^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\phi}_u^{(t)}) - F(\mathbf{C}^{(t+1)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\phi}_u^{(t)}) \geq \frac{\kappa_1}{2} \|\mathbf{C}^{(t+1)} - \mathbf{C}^{(t)}\|_F^2,$$

where $\kappa_1 = 2 - L_1 - \gamma_1$. As long as $\kappa_1 \geq 0$, the monotone descending property of the $\mathbf{C}$-step is guaranteed.

Similarly, we can investigate the $\boldsymbol{\beta}$-step. For any $\boldsymbol{\Delta}_2 \in \mathbb{R}^{(p_z+1)\times q}$,

$$H(\boldsymbol{\beta}^{(t+1)} + \boldsymbol{\Delta}_2; \mathbf{C}^{(t+1)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\phi}_u^{(t)}) - H(\boldsymbol{\beta}^{(t+1)}; \mathbf{C}^{(t+1)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\phi}_u^{(t)}) \geq \frac{1}{2} \|\boldsymbol{\Delta}_2\|_F^2,$$

by the triangular inequality. Based on a Taylor expansion, we get

$$\begin{aligned} &\sum_{(i,k)\in\Omega} \frac{1}{a_k(\phi_k^{(t)})} \{b_k(\mathbf{x}_i^\top \mathbf{c}_k^{(t+1)} + \mathbf{z}_i^\top \boldsymbol{\beta}_k^{(t+1)}) - b_k(\mathbf{x}_i^\top \mathbf{c}_k^{(t+1)} + \mathbf{z}_i^\top \boldsymbol{\beta}_k^{(t)})\} \\ &\quad = \sum_{(i,k)\in\Omega} \frac{1}{a_k(\phi_k^{(t)})} \{g_k^{-1}(\mathbf{x}_i^\top \mathbf{c}_k^{(t+1)} + \mathbf{z}_i^\top \boldsymbol{\beta}_k^{(t)})(\mathbf{z}_i^\top \boldsymbol{\beta}_k^{(t+1)} - \mathbf{z}_i^\top \boldsymbol{\beta}_k^{(t)}) \\ &\qquad + \frac{1}{2} b_k''(\mathbf{x}_i^\top \mathbf{c}_k^{(t+1)} + \mathbf{z}_i^\top \boldsymbol{\zeta}_k^{(t+1)})(\mathbf{z}_i^\top \boldsymbol{\beta}_k^{(t+1)} - \mathbf{z}_i^\top \boldsymbol{\beta}_k^{(t)})^2\}, \end{aligned}$$

where $\boldsymbol{\zeta}_k^{(t+1)} \in \{a\boldsymbol{\beta}_k^{(t)} + (1-a)\boldsymbol{\beta}_k^{(t+1)} : a \in (0,1)\}$. It follows that

$$H(\boldsymbol{\beta}^{(t+1)}; \mathbf{C}^{(t+1)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\phi}_u^{(t)}) \leq H(\boldsymbol{\beta}^{(t)}; \mathbf{C}^{(t+1)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\phi}_u^{(t)}) - \frac{1}{2}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_F^2 = F(\mathbf{C}^{(t+1)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\phi}_u^{(t)}) - \frac{1}{2}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_F^2.$$

Therefore,

$$\begin{aligned} &F(\mathbf{C}^{(t+1)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\phi}_u^{(t)}) - F(\mathbf{C}^{(t+1)}, \boldsymbol{\beta}^{(t+1)}, \boldsymbol{\phi}_u^{(t)}) \\ &\quad \geq \|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_F^2 - \frac{1}{2} \sum_{(i,k)\in\Omega} \frac{1}{a_k(\phi_k^{(t)})} b_k''(\mathbf{x}_i^\top \mathbf{c}_k^{(t+1)} + \mathbf{z}_i^\top \boldsymbol{\zeta}_k^{(t+1)})(\mathbf{z}_i^\top \boldsymbol{\beta}_k^{(t+1)} - \mathbf{z}_i^\top \boldsymbol{\beta}_k^{(t)})^2 \\ &\quad \geq \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{q} \left\{ 2 - \frac{1}{a_k(\phi_k^{(t)})} |b_k''(\mathbf{x}_i^\top \mathbf{c}_k^{(t+1)} + \mathbf{z}_i^\top \boldsymbol{\zeta}_k^{(t+1)})| \right\} (\mathbf{z}_i^\top \boldsymbol{\beta}_k^{(t+1)} - \mathbf{z}_i^\top \boldsymbol{\beta}_k^{(t)})^2. \end{aligned}$$

Now, let $\mathcal{B}_k = \{(\mathbf{c}_k^{(t+1)}, a\boldsymbol{\beta}_k^{(t)} + (1-a)\boldsymbol{\beta}_k^{(t+1)}, \phi_k^{(t)}) : a \in (0,1), t \in \{1,2,\ldots\}\}$, and

$$\gamma_2 = \max_{k \in \{1,\ldots,q\}} \sup_{(\boldsymbol{\xi}_k, \boldsymbol{\zeta}_k, \delta_k) \in \mathcal{B}_k} \|\mathbf{I}(\boldsymbol{\xi}_k, \boldsymbol{\zeta}_k, \delta_k; \mathbf{Z})\|_2.$$

Then

$$F(\mathbf{C}^{(t+1)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\phi}_u^{(t)}) - F(\mathbf{C}^{(t+1)}, \boldsymbol{\beta}^{(t+1)}, \boldsymbol{\phi}_u^{(t)}) \geq \frac{\kappa_2}{2} \|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|_F^2,$$

where $\kappa_2 = 2 - \gamma_2$. Finally, the unknown dispersion parameters are estimated based on maximizing the log-likelihood, so it is guaranteed to non-increase the objective function.

References

[1] T.W. Anderson, Estimating linear restrictions on regression coefficients for multivariate normal distributions, Ann. Math. Stat. 22 (1951) 327–351.

[2] A. Banerjee, S. Merugu, I.S. Dhillon, J. Ghosh, Clustering with Bregman divergences, J. Mach. Learn. Res. 6 (2005) 1705–1749.

[3] L.M. Bregman, The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming, USSR Comput. Math. Math. Phys. 7 (1967) 200–217.

[4] F. Bunea, Y. She, M.H. Wegkamp, Optimal selection of reduced rank estimators of high-dimensional matrices, Ann. Statist. 39 (2011) 1282–1309.

[5] F. Bunea, Y. She, M.H. Wegkamp, Joint variable and rank selection for parsimonious estimation of high dimensional matrices, Ann. Statist. 40 (2012) 2359–2388.
[6] J.-F. Cai, E.J. Candès, Z. Shen, A singular value thresholding algorithm for matrix completion, SIAM J. Optim. 20 (2010) 1956–1982.
[7] E.J. Candès, B. Recht, Exact matrix completion via convex optimization, Found. Comput. Math. 9 (2009) 717.
[8] K. Chen, R. Aseltine, Using hospitalization and mortality data to target suicide prevention activities: A demonstration from Connecticut, J. Adolesc. Health 61 (2017) 192–197.
[9] K. Chen, K.-S. Chan, N.C. Stenseth, Reduced rank stochastic regression with a sparse singular value decomposition, J. R. Stat. Soc. Ser. B 74 (2012) 203–221.
[10] K. Chen, H. Dong, K.-S. Chan, Reduced rank regression via adaptive nuclear norm penalization, Biometrika 100 (2013) 901–920.
[11] L. Chen, J.Z. Huang, Sparse reduced-rank regression for simultaneous dimension reduction and variable selection, J. Amer. Statist. Assoc. 107 (2012) 1533–1545.
[12] S. Choi, E.A. Hoffman, S.E. Wenzel, M. Castro, S.B. Fain, N.N. Jarjour, M.L. Schiebler, K. Chen, C.-L. Lin, Quantitative assessment of multiscale structural and functional alterations in asthmatic populations, J. Appl. Physiol. 118 (2015) 1286–1298.
[13] M. Collins, S. Dasgupta, R.E. Schapire, A generalization of principal components analysis to the exponential family, Adv. Neural Inf. Process. Syst. (2002) 617–624.
[14] R.D. Cook, L. Forzani, X. Zhang, Envelopes and reduced-rank regression, Biometrika 102 (2015) 439–456.
[15] R.D. Cook, S. Weisberg, Sliced inverse regression for dimension reduction: Comment, J. Amer. Statist. Assoc. 86 (1991) 328–332.
[16] D.R. Cox, N. Wermuth, Response models for mixed binary and quantitative variables, Biometrika 79 (1992) 441–461.
[17] D.B. Dunson, Bayesian latent variable models for clustered mixed outcomes, J. R. Stat. Soc. Ser. B 62 (2000) 355–366.
[18] Y. Fan, P. Lafaye de Micheaux, S. Penev, D. Salopek, Multivariate nonparametric test of independence, J. Multivariate Anal. 153 (2017) 189–210.
[19] J. Fan, J. Lv, A selective overview of variable selection in high dimensional feature space, Statist. Sinica 20 (2010) 101–148.
[20] Y. Fan, J. Lv, Asymptotic equivalence of regularization methods in thresholded parameter space, J. Amer. Statist. Assoc. 108 (2013) 1044–1061.
[21] Y. Fan, J. Lv, Asymptotic properties for combined ℓ_1 and concave regularization, Biometrika 101 (2014) 57–70.
[22] G.M. Fitzmaurice, N.M. Laird, Regression models for a bivariate discrete and continuous outcome with clustering, J. Amer. Statist. Assoc. 90 (1995) 845–852.
[23] A. Hammar, G. Ardal, Cognitive functioning in major depression: A summary, Front. Hum. Neurosci. 3 (2009) 26.
[24] A.E. Hoerl, R.W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, Technometrics 12 (1970) 55–67.
[25] B. Jorgensen, Exponential dispersion models, J. R. Stat. Soc. Ser. B 49 (1987) 127–162.
[26] O. Klopp, Noisy low-rank matrix completion with general sampling distribution, Bernoulli 20 (2014) 282–303.
[27] V. Koltchinskii, K. Lounici, A. Tsybakov, Nuclear norm penalization and optimal rates for noisy low rank matrix completion, Ann. Statist. 39 (2011) 2302–2329.
[28] J. Lafond, Low rank matrix completion with exponential family noise, J. Mach. Learn. Res.: Workshop Conf. Proc. 40 (2015) 1224–1243.
[29] K.-C. Li, Sliced inverse regression for dimension reduction, J. Amer. Statist. Assoc. 86 (1991) 316–327.
[30] B. Li, S. Wen, L. Zhu, On a projective resampling method for dimension reduction with multivariate responses, J. Amer. Statist. Assoc. 103 (2008) 1177–1186.
[31] L. Li, X. Zhang, Parsimonious tensor response regression, J. Amer. Statist. Assoc. 112 (2017) 1131–1146.
[32] K.-Y. Liang, S.L. Zeger, Longitudinal data analysis using generalized linear models, Biometrika 73 (1986) 13–22.
[33] Z. Lu, R.D.C. Monteiro, M. Yuan, Convex optimization methods for dimension reduction and coefficient estimation in multivariate linear regression, Math. Program. 131 (2012) 163–194.
[34] A. Mukherjee, K. Chen, N. Wang, J. Zhu, On the degrees of freedom of reduced-rank estimators in multivariate regression, Biometrika 102 (2015) 457–477.
[35] A. Mukherjee, J. Zhu, Reduced rank ridge regression and its kernel extensions, Stat. Anal. Data Min. 4 (2011) 612–622.
[36] S. Negahban, M.J. Wainwright, Estimation of (near) low-rank matrices with noise and high-dimensional scaling, Ann. Statist. 39 (2011) 1069–1097.
[37] J. Peng, J. Zhu, A. Bergamaschi, W. Han, D.-Y. Noh, J.R. Pollack, P. Wang, Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer, Ann. Appl. Stat. 4 (2010) 53–77.
[38] R.L. Prentice, L.P. Zhao, Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses, Biometrics 47 (1991) 825–839.
[39] G.C. Reinsel, P. Velu, Multivariate Reduced-Rank Regression: Theory and Applications, Springer, New York, 1998.
[40] A.J. Rothman, E. Levina, J. Zhu, Sparse multivariate regression with covariance estimation, J. Comput. Graph. Statist. 19 (2010) 947–962.
[41] M.D. Sammel, L.M. Ryan, J.M. Legler, Latent variable models for mixed discrete and continuous outcomes, J. R. Stat. Soc. Ser. B 59 (1997) 667–678.
[42] Y. She, Reduced rank vector generalized linear models for feature extraction, Stat. Interface 6 (2013) 197–209.
[43] Y. She, K. Chen, Robust reduced-rank regression, Biometrika 104 (2017) 633–647.
[44] X.-K. Song, P.X.-K. Song, Correlated Data Analysis: Modeling, Analytics, Applications, Springer, New York, 2007.
[45] D.C. Stanziano, M. Whitehurst, P. Graham, B.A. Roos, A review of selected longitudinal studies on aging: Past findings and future directions, J. Amer. Geriat. Soc. 58 (2010) S292–S297.
[46] M. Stone, Cross-validation and multinomial prediction, Biometrika 61 (1974) 509–515.
[47] G.J. Székely, M.L. Rizzo, N.K. Bakirov, Measuring and testing dependence by correlation of distances, Ann. Statist. 35 (2007) 2769–2794.
[48] M. Taha Bahadori, Z. Zheng, Y. Liu, J. Lv, Scalable Interpretable Multi-Response Regression via SEED. ArXiv e-prints, 2016.
[49] M. Udell, C. Horn, R. Zadeh, S. Boyd, Generalized low rank models, Found. Trends® Mach. Learn. 9 (2016) 1–118.
[50] T.W. Yee, T.J. Hastie, Reduced-rank vector generalized linear models, Stat. Model. 3 (2003) 15–41.
[51] M. Yuan, A. Ekici, Z. Lu, R. Monteiro, Dimension reduction and coefficient estimation in multivariate linear regression, J. R. Stat. Soc. Ser. B 69 (2007) 329–346.
[52] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, J. R. Stat. Soc. Ser. B 68 (2006) 49–67.
[53] L.P. Zhao, R.L. Prentice, S.G. Self, Multivariate mean parameter estimation by using a partly exponential model, J. R. Stat. Soc. Ser. B (1992) 805–811.
[54] H. Zhou, L. Li, Regularized matrix regression, J. R. Stat. Soc. Ser. B 76 (2014) 463–483.
[55] H. Zou, The adaptive lasso and its oracle properties, J. Amer. Statist. Assoc. 101 (2006) 1418–1429.