Statistica Sinica Preprint No: SS-2020-0401	
Title	Penalized Regression for Multiple Types of Many Features With Missing Data
Manuscript ID	SS-2020-0401
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202020.0401
Complete List of Authors	Kin Yau Wong, Donglin Zeng and Danyu Lin
Corresponding Author	Kin Yau Wong
E-mail	kin-yau.wong@polyu.edu.hk

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

PENALIZED REGRESSION FOR MULTIPLE TYPES OF MANY FEATURES WITH MISSING DATA

Kin Yau Wong[1], Donglin Zeng[2], and D. Y. Lin[2]

[1]*The Hong Kong Polytechnic University and* [2]*The University of North Carolina at Chapel Hill*

Abstract: Recent technological advances have made it possible to measure multiple types of many features in biomedical studies. However, some data types or features may not be measured for all study subjects because of cost or other constraints. We use a latent variable model to characterize the relationships across and within data types and to infer missing values from observed data. We develop a penalized-likelihood approach for variable selection and parameter estimation and devise an efficient expectation-maximization algorithm to implement our approach. We establish the asymptotic properties of the proposed estimators when the number of features increases at a polynomial rate of the sample size. Finally, we demonstrate the usefulness of the proposed methods using extensive simulation studies and provide an application to a motivating multi-platform genomics study.

Key words and phrases: Adaptive lasso, Factor models, Integrative analysis, Multi-modality data, Multi-platform genomics studies, Penalized regression

Corresponding author: Kin Yau Wong, Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong. E-mail: kin-yau.wong@polyu.edu.hk.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

1. Introduction

Modern biomedical studies often collect multiple types of data, or multi-modality data, on a large number of subjects. It is desirable to integrate such data because different modalities play unique roles in complex biological systems. For example, in the study of Alzheimer's disease, the integration of data on magnetic resonance imaging, positron emission tomography, and cerebrospinal fluid can yield more accurate disease classification (Zhang, Shen, and Alzheimer's Disease Neuroimaging Initiative, 2012). In cancer research, different types of genomics data, such as DNA alterations, RNA expressions, and protein expressions, can be integrated to identify disease subtypes and predict patient survival (Shen, Olshen, and Ladanyi, 2009; Wang et al., 2012; Hoadley et al., 2014; Wong et al., 2019).

Owing to cost or other constraints, certain features may not be measured on all study subjects. For example, in The Cancer Genome Atlas (TCGA) (https://cancergenome.nih.gov/), data on multiple types of genomic features, including DNA alterations, methylation profiles, and the expressions of RNA and protein, were collected for over 10,000 patients with 33 cancer types. For a substantial number of the patients, however, data on protein expressions were not generated. As another example, in the Trans-Omics for Precision Medicine (TOPMed) program (https://www.nhlbi.nih.gov/research/resources/nhlbi-precision-medicine-initiative/topmed), whole-genome sequencing data will be available for hundreds of thousands of subjects. However, other types of genomics data, such as RNA expressions, methylation profiles, and metabolites, will be available for only a few thousand subjects through ancillary studies of specific diseases.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

It is highly desirable to identify a small subset of features that are associated with the outcome of interest and to estimate the effects of these features. To perform variable selection and estimation with missing data, one may first produce a complete data set, and then apply conventional penalized regression methods (Tibshirani, 1996; Fan and Li, 2001; Zou, 2006) to the complete data set. Complete data may be obtained by deleting entries with missing data, mean imputation, and nearest-neighbor imputation (Troyanskaya et al., 2001). Recently, Cai, Cai, and Zhang (2016) proposed an imputation method specific to multi-modality data by assuming that the (complete) feature matrix is approximately low rank; the method is only applicable to a blockwise missing-data pattern, where a data type is either entirely missing or entirely observed on a subject. In general, the two-step approach to variable selection with missing data is inefficient, because it discards available data and ignores associations between the observed and missing variables. Also, the two-step approach yields inconsistent estimators when the data are not missing completely at random. To accommodate the missing-at-random mechanism, Ibrahim, Zhu, and Tang (2008), Garcia, Ibrahim, and Zhu (2010), and Jiang, Nguyen, and Rao (2015) proposed modeling the variables with missing values and performing variable selection using information-criterion or penalization methods. However, these approaches are intractable when there are many variables with missing values, as in our case.

Regression analysis on large, multi-modality data sets with missing values is highly challenging for two reasons. First, because different types of features tend to be correlated, efficient methods ought to leverage their relationships. However, it is difficult to formulate or estimate the intricate relationships between different types of many features. Second, in the presence of missing data, a tractable objective function for estimation is often

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

unavailable; for instance, the likelihood function would generally involve integration over many variables and not have a closed form.

To address the aforementioned challenges, we propose a penalized-likelihood approach in which the likelihood involves both an outcome model and a latent factor model for the potentially missing features. The factor model uses a small set of latent factors to explain the associations between features across or within individual data types, effectively reducing the dimensionality of the data. In multi-platform genomics studies, the latent factors can be interpreted as unobserved biological processes that govern the activities of different genomic features. This kind of model has been successfully used to combine multiple types of genomics data in order to understand the interactions between different types of features, recover personal genomics characteristics of cancer patients, and discover cancer subtypes (Shen, Olshen, and Ladanyi, 2009; Shen, Wang, and Mo, 2013; Lock et al., 2013).

Because the observed-data likelihood involves integration over the features with missing values, direct maximization of the (penalized) likelihood is computationally intensive or even infeasible when the number of features is moderately large. To efficiently compute the penalized estimators, we develop an expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977) that, by using the low-dimensional structure of the latent factor model, involves only low-dimensional integration. The algorithm is applicable to general missing-data patterns with a large number of features.

Because the likelihood involves the latent factor model, the total number of nonzero parameters is larger than the number of features. As a result, estimation consistency cannot be established under conventional high-dimensional settings, where the number of

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

features is larger than the sample size. In fact, it is highly challenging to establish the estimation and selection consistency of our penalized estimators, even when the number of features is smaller than but diverges with the sample size. In existing works on large latent factor models, estimation is based on the principal components analysis (Bai, 2003; Fan, Liao, and Mincheva, 2013; Fan, Liu, and Wang, 2018) or the maximization of the likelihood (Bai and Li, 2012; Bai and Liao, 2016). In those cases, the theoretical developments rely heavily on the specific closed-form expressions of the estimators or the likelihood. In our setting, variables may be missing, and the latent factor model is only a part of the full likelihood, which, in general, does not have a closed-form expression. In addition, proofs for the estimation and selection consistency of penalized regression methods for complete data (Fan and Peng, 2004; Fan and Lv, 2011) are not applicable, because an essential assumption about the lower bound of the eigenvalues of the information matrix does not hold for the latent factor model.

The rest of this paper is structured as follows. In Section 2, we formulate the model and define the maximum penalized-likelihood estimator. In Section 3, we describe the numerical implementation of the proposed methods. In Section 4, we present the asymptotic properties of the penalized estimators. In Section 5, we report the results from our simulation studies. In Section 6, we provide an application to a TCGA multi-platform genomics data set. We conclude the paper in Section 7, and relegate the theoretical details to the Appendix.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

2. Model, likelihood, and penalized estimation

Let Y be an outcome variable, $\boldsymbol{X}$ be a vector of covariates, and $(\boldsymbol{S}^{(1)}, \ldots, \boldsymbol{S}^{(K)})$ be K types of potentially missing covariates, for some $K \geq 1$. Let $\boldsymbol{S} = (\boldsymbol{S}^{(1)\mathrm{T}}, \ldots, \boldsymbol{S}^{(K)\mathrm{T}})^{\mathrm{T}}$. Suppose that the dimension of $\boldsymbol{X}$ is fixed, whereas the dimension of $\boldsymbol{S}$, denoted by p_n, may change with the sample size n. We specify the following models for Y and $\boldsymbol{S}$:

$$
\begin{aligned}
Y \mid (\boldsymbol{X}, \boldsymbol{S}) &\sim f(\cdot; \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{S}, \boldsymbol{\xi}), \\
\boldsymbol{S}^{(k)} &= \boldsymbol{\Gamma}^{(k)}\boldsymbol{X} + \boldsymbol{\Psi}^{(0,k)}\boldsymbol{U}^{(0)} + \boldsymbol{\Psi}^{(k)}\boldsymbol{U}^{(k)} + \boldsymbol{\epsilon}^{(k)} \qquad \text{for } k = 1, \ldots, K,
\end{aligned}
$$

where f is a parametric density function, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are vectors of regression parameters, $\boldsymbol{\xi}$ is a vector of low-dimensional nuisance parameters, $\boldsymbol{U}^{(k)}$ is a vector of multivariate standard-normal latent variables with dimension r_k ($r_k \geq 0$), for $k = 0, \ldots, K$, $(\boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Psi}^{(0,k)}, \boldsymbol{\Psi}^{(k)})_{k=1,\ldots,K}$ are matrices of regression parameters, and $(\boldsymbol{\epsilon}^{(1)}, \ldots, \boldsymbol{\epsilon}^{(K)})$ are zero-mean normal variables with independent components. The variables $(\boldsymbol{U}^{(0)}, \ldots, \boldsymbol{U}^{(K)}, \boldsymbol{\epsilon}^{(1)}, \ldots, \boldsymbol{\epsilon}^{(K)})$ are mutually independent. The numbers of latent variables $(r_0, \ldots, r_K)$ are chosen to be much smaller than the dimension of each type of features. To ensure model identifiability, we set $\psi_{jl}^{(0,1)} = 0$ and $\psi_{jl}^{(k)} = 0$, for $l > j$ and $k = 1, \ldots, K$, where $\psi_{jl}^{(0,1)}$ and $\psi_{jl}^{(k)}$ are the (j,l)th elements of $\boldsymbol{\Psi}^{(0,1)}$ and $\boldsymbol{\Psi}^{(k)}$, respectively. In addition, we assume that $\psi_{jj}^{(0,1)} > 0$ for $j = 1, \ldots, r_0$, and $\psi_{jj}^{(k)} > 0$ for $j = 1, \ldots, r_k$ and $k = 1, \ldots, K$. These conditions are analogous to condition (IC5) of Bai and Li (2012). They are satisfied if the first r_k components of the kth feature type depend on all corresponding type-specific latent variables $\boldsymbol{U}^{(k)}$, the first r_0 components of the first feature type depend on all common latent variables $\boldsymbol{U}^{(0)}$, and the corresponding vectors of factor loadings are linearly independent. In this case, the latent variables can be transformed to yield the desired

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

structure for the factor loading matrices. If these conditions are in doubt, we may refit the model under a different ordering of features. The model of Y, hereafter referred to as the outcome model, includes many common models, such as the linear and logistic regression models, as special cases. The model of $\boldsymbol{S}$ is a latent factor model, which assumes that $\boldsymbol{S}$ (conditional on $\boldsymbol{X}$) follows a multivariate normal distribution, and the associations between the multi-modality features are induced by a small set of unobserved latent factors $\boldsymbol{U} \equiv (\boldsymbol{U}^{(0)\mathrm{T}}, \ldots, \boldsymbol{U}^{(K)\mathrm{T}})^{\mathrm{T}}$.

The factor model captures the associations between features across different data types, as well as within individual data types. The set of latent variables $\boldsymbol{U}^{(0)}$ is shared among all data types and induces associations across all features. For $k = 1, \ldots, K$, the set of latent variables $\boldsymbol{U}^{(k)}$ is shared only among components of $\boldsymbol{S}^{(k)}$, and captures the associations between features of this data type that are not explained by $\boldsymbol{U}^{(0)}$. The factor model is plausible for many applications in which the features of individual or multiple types share common sources of variability. For example, in multi-platform cancer genomics studies, different types of genomic features are commonly affected by major biological processes, such as growth suppressor evasion and cell death resistance (Hanahan and Weinberg, 2011); associations induced by such processes can be captured by $\boldsymbol{U}^{(0)}$. In contrast, some biological processes, such as miRNA regulation, may alter the expression of genes with no effect on other types of features, such as mutations; associations induced by such processes can be captured by the type-specific latent variables $\boldsymbol{U}^{(k)}$ $(k = 1, \ldots, K)$.

We allow each component of $\boldsymbol{S} \equiv (S_1, \ldots, S_{p_n})^{\mathrm{T}}$ to be missing and use M_j to indicate, by the values one versus zero, whether S_j is observed or missing $(j = 1, \ldots, p_n)$, respectively. We assume that $\boldsymbol{S}$ is missing at random, such that $\boldsymbol{M} \equiv (M_1, \ldots, M_{p_n})^{\mathrm{T}}$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

is independent of $\boldsymbol{S}$ conditional on $(Y, \boldsymbol{X}, \overline{\boldsymbol{S}})$, where $\overline{\boldsymbol{S}} = \{S_j : P(M_j = 1) = 1\}$. This assumption holds when missing data are introduced by design, where subjects with specific values of $(Y, \boldsymbol{X}, \overline{\boldsymbol{S}})$ are selected for measurements of components of $\boldsymbol{S}$ (not included in $\overline{\boldsymbol{S}}$). The assumption also holds when missing data arise from random technical errors in the data-collection process that are independent of the data.

For a random sample of size n, the observed data consist of $(Y_i, \boldsymbol{X}_i, \boldsymbol{M}_i, \boldsymbol{M}_i \circ \boldsymbol{S}_i)$ $(i = 1, \ldots, n)$, where $\circ$ denotes componentwise multiplication. Let $r = \sum_{k=0}^{K} r_k$, $\boldsymbol{\Gamma} = (\boldsymbol{\Gamma}^{(1)\mathrm{T}}, \ldots, \boldsymbol{\Gamma}^{(K)\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\Psi}$ be a $(p_n \times r)$ matrix with

$$\boldsymbol{\Psi} = \begin{pmatrix} \boldsymbol{\Psi}^{(0,1)} & \boldsymbol{\Psi}^{(1)} & \mathbf{0} & \cdots & \mathbf{0} \\ \boldsymbol{\Psi}^{(0,2)} & \mathbf{0} & \boldsymbol{\Psi}^{(2)} & \cdots & \mathbf{0} \\ \vdots & \vdots & & \ddots & \vdots \\ \boldsymbol{\Psi}^{(0,K)} & \mathbf{0} & \cdots & \mathbf{0} & \boldsymbol{\Psi}^{(K)} \end{pmatrix}, \tag{2.1}$$

and $\boldsymbol{\Sigma}$ be a $(p_n \times p_n)$ diagonal matrix with the diagonal elements being the variances of the components of $(\boldsymbol{\epsilon}^{(1)}, \ldots, \boldsymbol{\epsilon}^{(K)})$. Let $\boldsymbol{\theta} \equiv (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\Gamma}, \boldsymbol{\Psi}, \boldsymbol{\Sigma})$ denote the collection of all parameters. The observed-data log-likelihood function concerning $\boldsymbol{\theta}$ is

$$\ell_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \int f(Y_i; \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{X}_i + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{S}_i, \boldsymbol{\xi}) \phi(\boldsymbol{S}_i; \boldsymbol{\Gamma} \boldsymbol{X}_i, \boldsymbol{\Psi}\boldsymbol{\Psi}^{\mathrm{T}} + \boldsymbol{\Sigma}) \, \mathrm{d}\boldsymbol{S}_i^{(M)}, \tag{2.2}$$

where $\boldsymbol{S}_i^{(M)}$ is the vector of the missing components of $\boldsymbol{S}_i$, and $\phi(\cdot; \boldsymbol{\mu}, \boldsymbol{\Omega})$ is the density of the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Omega}$. We propose estimating $\boldsymbol{\theta}$ using maximum penalized-likelihood estimation with an adaptive lasso (Zou, 2006) penalty on $\boldsymbol{\beta}$. Specifically, the penalized estimator $\widehat{\boldsymbol{\theta}} \equiv (\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\xi}}, \widehat{\boldsymbol{\Gamma}}, \widehat{\boldsymbol{\Psi}}, \widehat{\boldsymbol{\Sigma}})$ maximizes

$$p\ell_n(\boldsymbol{\theta}) \equiv \ell_n(\boldsymbol{\theta}) - n\lambda_n \sum_{j=1}^{p_n} w_j |\beta_j|,$$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

where λ_n is a tuning parameter, and $w_j \equiv |\widetilde{\beta}_j|^{-1}$ is a weight term derived from an initial estimator $\widetilde{\beta}_j$ $(j = 1, \ldots, p_n)$.

In general, the likelihood involves the conditional distribution of Y given the observed components of $\boldsymbol{S}$ (and $\boldsymbol{X}$) and the distribution of the observed components of $\boldsymbol{S}$. When the data are complete, the conditional distribution of Y involves $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi})$ only, and the distribution of $\boldsymbol{S}$ involves $(\boldsymbol{\Gamma}, \boldsymbol{\Psi}, \boldsymbol{\Sigma})$ only, such that the two sets of parameters can be estimated separately. With missing data, however, the conditional distribution of Y given the observed components of $\boldsymbol{S}$ involves $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi})$ and functions of $(\boldsymbol{\Gamma}, \boldsymbol{\Psi}, \boldsymbol{\Sigma})$ that capture the relationship between the missing and observed components of $\boldsymbol{S}$. Therefore, valid estimation and variable selection for the outcome model ought to properly account for the relationships between different components of $\boldsymbol{S}$. Mean imputation completely ignores these relationships, and single imputation based on the observed components of $\boldsymbol{S}$ alone may be biased when the missing-data mechanism depends on $(Y, \boldsymbol{X})$. Both approaches may yield inefficient or inconsistent estimation of the outcome model.

To obtain the initial estimators $\widetilde{\beta}_j$ $(j = 1, \ldots, p_n)$, one may maximize $\ell_n(\cdot)$ with an L_1 or L_2 penalty on $\boldsymbol{\beta}$. However, this approach involves an extra step of tuning parameter selection and is computationally intensive, owing to the missing data. An alternative approach is to fit a "marginal" regression model of Y against $(\boldsymbol{X}, S_j)$ for each j, and use the regression parameter estimator for S_j as the initial estimator $\widetilde{\beta}_j$ $(j = 1, \ldots, p_n)$. In each model, we assume that S_j follows a linear regression model with covariates $\boldsymbol{X}$. Involving only a single incomplete independent variable, the model can be estimated easily using the EM algorithm. We adopt the marginal approach because it is computationally efficient and does not require tuning. We expect the marginal approach to perform well

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

when the marginal effects of $\boldsymbol{S}$ have a sparsity structure similar to that of the conditional effects.

3. Computation of the penalized estimators

It is convenient to introduce the notation $\mathcal{M}_i = \{j : M_{ij} = 0\}$, and $\mathcal{M}_i^{\mathrm{C}} = \{1, \ldots, p_n\} \backslash \mathcal{M}_i$. In the remainder of the paper, b_{ij} denotes the jth component of the vector $\boldsymbol{b}_i$. Let $\boldsymbol{S}_i^{(O)} = (\boldsymbol{S}_i)_{\mathcal{M}_i^{\mathrm{C}}}$, $\boldsymbol{\beta}_i^{(M)} = \boldsymbol{\beta}_{\mathcal{M}_i}$, $\boldsymbol{\beta}_i^{(O)} = \boldsymbol{\beta}_{\mathcal{M}_i^{\mathrm{C}}}$, $\boldsymbol{\epsilon}_i^{(M)} = (\boldsymbol{\epsilon}_i)_{\mathcal{M}_i}$, $\boldsymbol{\Gamma}_i^{(M)} = \boldsymbol{\Gamma}_{\mathcal{M}_i}$, and $\boldsymbol{\Psi}_i^{(M)} = \boldsymbol{\Psi}_{\mathcal{M}_i}$, where $\boldsymbol{b}_{\mathcal{A}}$ is a vector that consists of all components of $\boldsymbol{b}$ indexed by $\mathcal{A}$, and $\boldsymbol{B}_{\mathcal{A}}$ is a matrix that consists of all rows of $\boldsymbol{B}$ indexed by $\mathcal{A}$. By the definition of the factor model, the likelihood function is proportional to

$$\prod_{i=1}^{n} \int f(Y_i; \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{X}_i + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{S}_i, \boldsymbol{\xi}) \prod_{j=1}^{p_n} \sigma_j^{-1} \exp\left\{ -\frac{(S_{ij} - \boldsymbol{\gamma}_j^{\mathrm{T}} \boldsymbol{X}_i - \boldsymbol{\psi}_j^{\mathrm{T}} \boldsymbol{U}_i)^2}{2\sigma_j^2} \right\} e^{-\frac{1}{2}\boldsymbol{U}_i^{\mathrm{T}} \boldsymbol{U}_i} \, \mathrm{d}(\boldsymbol{U}_i, \boldsymbol{S}_i^{(M)}),$$

where $\boldsymbol{\gamma}_j^{\mathrm{T}}$ and $\boldsymbol{\psi}_j^{\mathrm{T}}$ are the jth rows of $\boldsymbol{\Gamma}$ and $\boldsymbol{\Psi}$, respectively, and σ_j^2 is the jth diagonal element of $\boldsymbol{\Sigma}$. With $S_{ij} = \boldsymbol{\gamma}_j^{\mathrm{T}} \boldsymbol{X}_i + \boldsymbol{\psi}_j^{\mathrm{T}} \boldsymbol{U}_i + \epsilon_{ij}$, for $M_{ij} = 0$, the above expression becomes

$$\begin{aligned} &\prod_{i=1}^{n} \int f(Y_i; \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{X}_i + \boldsymbol{\beta}_i^{(O)\mathrm{T}} \boldsymbol{S}_i^{(O)} + \boldsymbol{\beta}_i^{(M)\mathrm{T}} \boldsymbol{\Gamma}_i^{(M)} \boldsymbol{X}_i + \boldsymbol{\beta}_i^{(M)\mathrm{T}} \boldsymbol{\Psi}_i^{(M)} \boldsymbol{U}_i + \boldsymbol{\beta}_i^{(M)\mathrm{T}} \boldsymbol{\epsilon}_i^{(M)}, \boldsymbol{\xi}) \\ &\times \prod_{j=1}^{p_n} \sigma_j^{-1} \exp\left[-\frac{1}{2\sigma_j^2} \left\{ M_{ij} (S_{ij} - \boldsymbol{\gamma}_j^{\mathrm{T}} \boldsymbol{X}_i - \boldsymbol{\psi}_j^{\mathrm{T}} \boldsymbol{U}_i)^2 + (1 - M_{ij}) \epsilon_{ij}^2 \right\} \right] e^{-\frac{1}{2}\boldsymbol{U}_i^{\mathrm{T}} \boldsymbol{U}_i} \, \mathrm{d}(\boldsymbol{U}_i, \boldsymbol{\epsilon}_i^{(M)}). \end{aligned} \tag{3.1}$$

To obtain the penalized estimators, we adopt an EM algorithm with $(\boldsymbol{U}_i, \boldsymbol{\epsilon}_i^{(M)})$ $(i = 1, \ldots, n)$ as missing data. The algorithm iterates between the E-step and M-step, described below, until convergence. In contrast to direct maximization of the log-likelihood function, the EM algorithm avoids inversion of large matrices and involves numerical integration of lower dimensions.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

In the E-step, we calculate the conditional expectation of functions of $(\boldsymbol{U}_i, \boldsymbol{\epsilon}_i^{(M)})$ that are involved in the M-step. Because all functions of $\boldsymbol{\epsilon}_i^{(M)}$ involved in the M-step are linear or quadratic, we need only calculate the conditional expectation of the functions of $\boldsymbol{U}_i$ and a one-dimensional linear transformation of $\boldsymbol{\epsilon}_i^{(M)}$. Let $p_i^{(M)}$ be the dimension of $\boldsymbol{\epsilon}_i^{(M)}$, $c_i = \Big\{ \sum_{j=1}^{p_i^{(M)}} (\beta_{ij}^{(M)} \sigma_{ij}^{(M)})^2 \Big\}^{1/2}$, and $\widetilde{\epsilon}_i = \boldsymbol{\beta}_i^{(M)\mathrm{T}} \boldsymbol{\epsilon}_i^{(M)}$. Because $\widetilde{\epsilon}_i$ is zero-mean normal with variance c_i and is independent of $\boldsymbol{U}_i$ and $\{\epsilon_{ij} : j \in \mathcal{M}_i^{\mathrm{C}}\}$, the joint density function of $(Y_i, \boldsymbol{S}_i, \boldsymbol{U}_i, \widetilde{\epsilon}_i)$ is proportional to

$$f(Y_i, \boldsymbol{S}_i, \boldsymbol{U}_i, \widetilde{\epsilon}_i; \boldsymbol{X}_i) \equiv f(Y_i; \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{X}_i + \boldsymbol{\beta}_i^{(O)\mathrm{T}} \boldsymbol{S}_i^{(O)} + \boldsymbol{\beta}_i^{(M)\mathrm{T}} \boldsymbol{\Gamma}_i^{(M)} \boldsymbol{X}_i + \boldsymbol{\beta}_i^{(M)\mathrm{T}} \boldsymbol{\Psi}_i^{(M)} \boldsymbol{U}_i + \widetilde{\epsilon}_i, \boldsymbol{\xi}) \times \exp\Big\{ - \sum_{j \in \mathcal{M}_i^{\mathrm{C}}} \frac{1}{2\sigma_j^2} (\boldsymbol{S}_j - \boldsymbol{\gamma}_j^{\mathrm{T}} \boldsymbol{X}_i - \boldsymbol{\psi}_j^{\mathrm{T}} \boldsymbol{U}_i)^2 - \frac{\widetilde{\epsilon}_i^2}{2c_i^2} - \frac{1}{2} \boldsymbol{U}_i^{\mathrm{T}} \boldsymbol{U}_i \Big\}.$$

The conditional expectation of any function g of $(\boldsymbol{U}_i, \widetilde{\epsilon}_i)$ given the observed data is

$$\mathcal{C}^{-1} \int g(\boldsymbol{U}_i, \widetilde{\epsilon}_i) f(Y_i, \boldsymbol{S}_i, \boldsymbol{U}_i, \widetilde{\epsilon}_i; \boldsymbol{X}_i) \, \mathrm{d}(\boldsymbol{U}_i, \widetilde{\epsilon}_i), \tag{3.2}$$

where $\mathcal{C}$ is equal to the above integral evaluated at $g(\cdot, \cdot) = 1$. In contrast to the $(r + p_i^{(M)})$-dimensional integration in (3.1), the integration in (3.2) is of dimension $(r + 1)$ only. To approximate (3.2), we extend the approach of Liu and Pierce (1994) to the multivariate setting, and write (3.2) as

$$\int w(\boldsymbol{\upsilon}) e^{-\frac{1}{2}(\boldsymbol{\upsilon} - \boldsymbol{\mu}_\upsilon)^{\mathrm{T}} \boldsymbol{H}_\upsilon^{-1} (\boldsymbol{\upsilon} - \boldsymbol{\mu}_\upsilon)} \, \mathrm{d}\boldsymbol{\upsilon}, \tag{3.3}$$

where $\boldsymbol{\upsilon} = (\boldsymbol{U}_i^{\mathrm{T}}, \widetilde{\epsilon}_i)^{\mathrm{T}}$, $w(\boldsymbol{\upsilon}) = \mathcal{C}^{-1} g(\boldsymbol{U}_i, \widetilde{\epsilon}_i) f(Y_i, \boldsymbol{S}_i, \boldsymbol{U}_i, \widetilde{\epsilon}_i; \boldsymbol{X}_i) e^{\frac{1}{2}(\boldsymbol{\upsilon} - \boldsymbol{\mu}_\upsilon)^{\mathrm{T}} \boldsymbol{H}_\upsilon^{-1} (\boldsymbol{\upsilon} - \boldsymbol{\mu}_\upsilon)}$, $\boldsymbol{\mu}_\upsilon$ is the maximizer of $f(Y_i, \boldsymbol{S}_i, \boldsymbol{U}_i, \widetilde{\epsilon}_i; \boldsymbol{X}_i)$ with respect to $(\boldsymbol{U}_i, \widetilde{\epsilon}_i)$, and $\boldsymbol{H}_\upsilon$ is the Hessian matrix of $-\log f(Y_i, \boldsymbol{S}_i, \boldsymbol{U}_i, \widetilde{\epsilon}_i; \boldsymbol{X}_i)$ with respect to $(\boldsymbol{U}_i, \widetilde{\epsilon}_i)$ evaluated at $\boldsymbol{\mu}_\upsilon$. Then, we can approximate (3.3) using the sparse-grid multivariate Gauss-Hermite quadrature (Heiss and Winschel, 2008). Unlike a conventional multivariate quadrature, where the number

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

of nodes increases exponentially with the dimension of integration under a fixed level of accuracy, the number of nodes under the sparse-grid quadrature increases only polynomially with the dimension. All functions of $\boldsymbol{\epsilon}_i^{(M)}$ involved in the M-step can be obtained from the first and second moments of $\widetilde{\epsilon}_i$. The relationship between the moments of $(\boldsymbol{U}_i, \boldsymbol{\epsilon}_i^{(M)})$ and the moments of $(\boldsymbol{U}_i, \widetilde{\epsilon}_i)$ is given in Appendix A.1.

To perform the M-step, we obtain a local quadratic approximation to the log-likelihood of the outcome model. For $i = 1, \ldots, n$, let $\widehat{\eta}_i = \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{X}_i + \boldsymbol{\beta}_i^{(O)\mathrm{T}} \boldsymbol{S}_i^{(O)} + \boldsymbol{\beta}_i^{(M)\mathrm{T}} \boldsymbol{\Gamma}_i^{(M)} \boldsymbol{X}_i + \boldsymbol{\beta}_i^{(M)\mathrm{T}} \boldsymbol{\Psi}_i^{(M)} \widehat{\mathrm{E}}(\boldsymbol{U}_i) + \widehat{\mathrm{E}}(\widetilde{\epsilon}_i)$, where $\widehat{\mathrm{E}}(\cdot)$ is the conditional expectation obtained from the E-step, and the parameters are evaluated at the estimators obtained from the previous M-step (or the initial estimators for the first iteration). By the Taylor series expansion of $\log f(Y_i; \eta, \boldsymbol{\xi})$ at $\eta = \widehat{\eta}_i$, we can approximate $\log f(Y_i; \eta, \boldsymbol{\xi})$ by $-(z_i - \eta)^2/(2s_i^2)$ up to a constant term, where $s_i^2 = -\partial^2 \log f(Y_i; \eta, \boldsymbol{\xi})/\partial \eta^2|_{\eta=\widehat{\eta}_i}$, and $z_i = \widehat{\eta}_i + s_i^{-2} \partial \log f(Y_i; \eta, \boldsymbol{\xi})/\partial \eta|_{\eta=\widehat{\eta}_i}$.

In the M-step, we first update the parameters in the latent factor model in a coordinate-wise fashion by maximizing

$$
\begin{aligned}
&-\sum_{i=1}^{n} \widehat{\mathrm{E}} \Bigg\{ \frac{M_{ij}}{2\sigma_j^2} \big(S_{ij} - \boldsymbol{\gamma}_j^{\mathrm{T}} \boldsymbol{X}_i - \boldsymbol{\psi}_j^{\mathrm{T}} \boldsymbol{U}_i\big)^2 \\
&\quad + \frac{1 - M_{ij}}{2s_i^2} \big(z_i - \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{X}_i - \boldsymbol{\beta}_i^{(O)\mathrm{T}} \boldsymbol{S}_i^{(O)} - \boldsymbol{\beta}_i^{(M)\mathrm{T}} \boldsymbol{\Gamma}_i^{(M)} \boldsymbol{X}_i - \boldsymbol{\beta}_i^{(M)\mathrm{T}} \boldsymbol{\Psi}_i^{(M)} \boldsymbol{U}_i - \widetilde{\epsilon}_i\big)^2 \Bigg\}
\end{aligned}
$$

with respect to $(\boldsymbol{\gamma}_j, \boldsymbol{\psi}_j, \sigma_j^2)$ in turn, for $j = 1, \ldots, p_n$, where the remaining parameters are fixed at the current estimators. Note that if S_j belongs to the kth data type ($j = 1, \ldots, p_n; k = 1, \ldots, K$), then only the components of $\boldsymbol{\psi}_j$ that correspond to $\boldsymbol{U}^{(0)}$ and $\boldsymbol{U}^{(k)}$ need to be updated; the remaining components are set to zero. Furthermore, under the identifiability conditions, the upper triangular elements of $\boldsymbol{\Psi}^{(0,1)}$ and $\boldsymbol{\Psi}^{(1)}, \ldots, \boldsymbol{\Psi}^{(K)}$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

are set to zero. Then, we update $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by maximizing

$$\sum_{i=1}^{n} \frac{z_i}{s_i^2}\{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}_i + \boldsymbol{\beta}^{\mathrm{T}}\widehat{\mathrm{E}}(\widetilde{\boldsymbol{S}}_i)\} - \frac{1}{2s_i^2}(\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}) \begin{pmatrix} \boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}} & \boldsymbol{X}_i\widehat{\mathrm{E}}(\widetilde{\boldsymbol{S}}_i^{\mathrm{T}}) \\ \widehat{\mathrm{E}}(\widetilde{\boldsymbol{S}}_i)\boldsymbol{X}_i^{\mathrm{T}} & \widehat{\mathrm{E}}(\widetilde{\boldsymbol{S}}_i\widetilde{\boldsymbol{S}}_i^{\mathrm{T}}) \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} - \lambda_n|\boldsymbol{w} \circ \boldsymbol{\beta}|, \tag{3.4}$$

where $\widetilde{\boldsymbol{S}}_i = \boldsymbol{M}_i \circ \boldsymbol{S}_i + (\mathbf{1} - \boldsymbol{M}_i) \circ (\widehat{\boldsymbol{\Gamma}}\boldsymbol{X}_i + \widehat{\boldsymbol{\Psi}}\boldsymbol{U}_i + \boldsymbol{\epsilon}_i)$, and $(\widehat{\boldsymbol{\Gamma}}, \widehat{\boldsymbol{\Psi}})$ are the current estimators of $(\boldsymbol{\Gamma}, \boldsymbol{\Psi})$. The estimators of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ can be computed efficiently using the coordinate-descent algorithm (Friedman, Hastie, and Tibshirani, 2010) for complete data. Finally, we update the estimator of $\boldsymbol{\xi}$ by maximizing the conditional expected log-likelihood with other parameters evaluated at the current estimators.

To obtain more stable estimators, we adopt the pathwise coordinate-descent approach of Friedman, Hastie, and Tibshirani (2010): instead of directly computing the penalized estimators at the selected value of λ_n, we perform the estimation for a sequence of decreasing values of λ_n up to the selected value. The sequence starts at $\lambda_{\max}$, the smallest value of λ_n under which all estimators of $\boldsymbol{\beta}$ are zero. In particular, $\lambda_{\max} = \max_j w_j^{-1}|n^{-1}\sum_{i=1}^{n} s_i^{-2}(z_i - \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}_i)\widehat{\mathrm{E}}(\widetilde{S}_{ij})|$, where $\boldsymbol{\alpha}$ and the parameters in s_i, z_i, and $\widehat{\mathrm{E}}(\cdot)$ are evaluated at the maximum likelihood estimators (MLE) under $\boldsymbol{\beta} = \mathbf{0}$. For the linear outcome model, the estimator of the error variance may vary greatly over different values of λ_n. Therefore, the estimators of $\boldsymbol{\beta}$ obtained from maximizing (3.4) may be unstable over different values of λ_n. To overcome this problem, we maximize (3.4) with s_i set to one in each M-step, and also set $s_i = 1$ in the calculation of $\lambda_{\max}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

4. Asymptotic properties of the penalized estimators

We partition $\boldsymbol{\beta}$ as $(\boldsymbol{\beta}_{\mathcal{S}}^{\mathrm{T}}, \boldsymbol{\beta}_{\mathcal{N}}^{\mathrm{T}})^{\mathrm{T}}$, such that $\boldsymbol{\beta}_{\mathcal{S}}$ is p_{1n}-dimensional and has a nonzero true value, and $\boldsymbol{\beta}_{\mathcal{N}}$ has a true value of $\mathbf{0}$. Write $\boldsymbol{w} = (w_1, \ldots, w_{p_n})^{\mathrm{T}}$, and partition $\boldsymbol{w} = (\boldsymbol{w}_{\mathcal{S}}^{\mathrm{T}}, \boldsymbol{w}_{\mathcal{N}}^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{S} = (\boldsymbol{S}_{\mathcal{S}}^{\mathrm{T}}, \boldsymbol{S}_{\mathcal{N}}^{\mathrm{T}})^{\mathrm{T}}$ to conform with the partitioning of $\boldsymbol{\beta}$. Assume that $p_n = O(n^{\kappa})$, for some positive $\kappa < 1/5$ and $p_{1n} = O(p_n)$.

Let $Z(\eta, \boldsymbol{\xi}) = \partial \log f(Y; \eta, \boldsymbol{\xi})/\partial \eta$, $\dot{\boldsymbol{\ell}}_{\alpha}^{(C)}(\boldsymbol{\theta}) = Z(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{S}, \boldsymbol{\xi})\boldsymbol{X}$, $\dot{\boldsymbol{\ell}}_{\beta_{\mathcal{S}}}^{(C)}(\boldsymbol{\theta}) = Z(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{S}, \boldsymbol{\xi})\boldsymbol{S}_{\mathcal{S}}$, $\dot{\boldsymbol{\ell}}_{\beta_{\mathcal{N}}}^{(C)}(\boldsymbol{\theta}) = Z(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{S}, \boldsymbol{\xi})\boldsymbol{S}_{\mathcal{N}}$, $\dot{\boldsymbol{\ell}}_{\xi}^{(C)}(\boldsymbol{\theta}) = \partial \log f(Y; \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{S}, \boldsymbol{\xi})/\partial \boldsymbol{\xi}$, $\dot{\boldsymbol{\ell}}_{\Gamma}^{(C)}(\boldsymbol{\theta}) = \boldsymbol{\Omega}^{-1}(\boldsymbol{S} - \boldsymbol{\Gamma}\boldsymbol{X})\boldsymbol{X}^{\mathrm{T}}$, $\dot{\boldsymbol{\ell}}_{\Psi}^{(C)}(\boldsymbol{\theta}) = \boldsymbol{\Omega}^{-1}\{(\boldsymbol{S} - \boldsymbol{\Gamma}\boldsymbol{X})(\boldsymbol{S} - \boldsymbol{\Gamma}\boldsymbol{X})^{\mathrm{T}} - \boldsymbol{\Omega}\}\boldsymbol{\Omega}^{-1}\boldsymbol{\Psi}$, and $\dot{\boldsymbol{\ell}}_{\Sigma}^{(C)}(\boldsymbol{\theta}) = \mathrm{diag}\big[\boldsymbol{\Omega}^{-1}\{(\boldsymbol{S} - \boldsymbol{\Gamma}\boldsymbol{X})(\boldsymbol{S} - \boldsymbol{\Gamma}\boldsymbol{X})^{\mathrm{T}} - \boldsymbol{\Omega}\}\boldsymbol{\Omega}^{-1}\big]$, where $\boldsymbol{\Omega} = \boldsymbol{\Psi}\boldsymbol{\Psi}^{\mathrm{T}} + \boldsymbol{\Sigma}$, and $\mathrm{diag}(\boldsymbol{D})$ is the diagonal matrix that consists of the diagonal elements of $\boldsymbol{D}$. Let

$$\dot{\boldsymbol{\ell}}_{\theta_{\mathcal{S}}}^{(C)}(\boldsymbol{\theta}) \equiv (\dot{\boldsymbol{\ell}}_{\alpha}^{(C)}(\boldsymbol{\theta})^{\mathrm{T}}, \dot{\boldsymbol{\ell}}_{\beta_{\mathcal{S}}}^{(C)}(\boldsymbol{\theta})^{\mathrm{T}}, \dot{\boldsymbol{\ell}}_{\xi}^{(C)}(\boldsymbol{\theta})^{\mathrm{T}}, \mathrm{vec}\{\dot{\boldsymbol{\ell}}_{\Gamma}^{(C)}(\boldsymbol{\theta})\}^{\mathrm{T}}, \mathrm{vec}\{\dot{\boldsymbol{\ell}}_{\Psi}^{(C)}(\boldsymbol{\theta})\}^{\mathrm{T}}, \mathrm{vecd}\{\dot{\boldsymbol{\ell}}_{\Sigma}^{(C)}(\boldsymbol{\theta})\}^{\mathrm{T}})^{\mathrm{T}}$$

be a vector of the score statistics for a subject with complete data, where $\mathrm{vec}(\boldsymbol{D})$ denotes the vector obtained from stacking the columns of $\boldsymbol{D}$, and $\mathrm{vecd}(\boldsymbol{D})$ denotes the vector of the diagonal elements of $\boldsymbol{D}$. Define $\boldsymbol{V}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi}) = \partial^2 \log f(\boldsymbol{Y}; \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{S}, \boldsymbol{\xi})$ $/\partial(\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\xi}^{\mathrm{T}})^{\mathrm{T}}\partial(\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\xi}^{\mathrm{T}})$ and $\boldsymbol{I}(\boldsymbol{\theta}) = \mathrm{E}\big[\mathrm{E}\{\dot{\boldsymbol{\ell}}_{\beta_{\mathcal{N}}}^{(C)}(\boldsymbol{\theta}) \mid \mathcal{O}\}\mathrm{E}\{\dot{\boldsymbol{\ell}}_{\theta_{\mathcal{S}}}^{(C)}(\boldsymbol{\theta})^{\mathrm{T}} \mid \mathcal{O}\}\big]$, where $\mathcal{O}$ denotes the observed data, which consist of $(Y, \boldsymbol{X})$ and a (random) subset of $\boldsymbol{S}$. Let $\boldsymbol{\beta}_0 \equiv (\boldsymbol{\beta}_{0\mathcal{S}}^{\mathrm{T}}, \boldsymbol{\beta}_{0\mathcal{N}}^{\mathrm{T}})^{\mathrm{T}} \equiv (\beta_{01}, \ldots, \beta_{0p_n})^{\mathrm{T}}$ denote the true value of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}_0 \equiv (\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0, \boldsymbol{\xi}_0, \boldsymbol{\Gamma}_0, \boldsymbol{\Psi}_0, \boldsymbol{\Sigma}_0)$ denote the true value of $\boldsymbol{\theta}$. For $k = 1, \ldots, K$, let $\boldsymbol{\Psi}_0^{(0,k)}$ and $\boldsymbol{\Psi}_0^{(k)}$ be the true values of $\boldsymbol{\Psi}^{(0,k)}$ and $\boldsymbol{\Psi}^{(k)}$, respectively. In the following, $\|\cdot\|$ denotes the L_2 norm for vectors or the Frobenius norm for matrices.

We impose the following conditions, some of which involve a generic, finite, and positive constant C.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

(C1) The vector of covariates $\boldsymbol{X}$ is bounded, and the eigenvalues of $\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})$ lie within (C^{-1}, C). Furthermore, each component of $(\dot{\boldsymbol{\ell}}_{\alpha}^{(C)}(\boldsymbol{\theta}_0), \dot{\boldsymbol{\ell}}_{\beta_{\mathcal{S}}}^{(C)}(\boldsymbol{\theta}_0), \dot{\boldsymbol{\ell}}_{\beta_{\mathcal{N}}}^{(C)}(\boldsymbol{\theta}_0), \dot{\boldsymbol{\ell}}_{\xi}^{(C)}(\boldsymbol{\theta}_0))$ has a finite second moment, and $\lambda_{\min}\{-\mathrm{E}\boldsymbol{V}(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0, \boldsymbol{\xi}_0)\} > C^{-1}$, where $\lambda_{\min}(\boldsymbol{D})$ is the smallest eigenvalue of $\boldsymbol{D}$. In addition, within a small neighborhood of $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0, \boldsymbol{\xi}_0)$ and for any element $v(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi})$ of $\boldsymbol{V}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi})$, $v(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi})$ is twice differentiable, with each component of its derivatives up to the second order uniformly bounded by a function of $(Y, \boldsymbol{X}, \boldsymbol{S})$ that has a finite second moment.

(C2) The probability $P(M_1 = 1, \ldots, M_{p_n} = 1 \mid Y, \boldsymbol{X}, \overline{\boldsymbol{S}}) > C^{-1}$ for almost surely all $(Y, \boldsymbol{X}, \overline{\boldsymbol{S}})$.

(C3) The initial estimators satisfy that $|\widetilde{\beta}_j|^{-1} = O_p(n^{\rho})$ for $j = 1, \ldots, p_{1n}$ and some $\rho \in [0, 1/2)$, and $|\widetilde{\beta}_j| = O_p(n^{-\tau})$ for $j = p_{1n} + 1, \ldots, p_n$ and some $\tau \in (\kappa, 1/2]$.

(C4) The tuning parameter λ_n satisfies $\lambda_n n^{1/2+\rho} \to 0$ as $n \to \infty$, where ρ is defined in condition (C3).

(C5) The parameters $(\boldsymbol{\Gamma}, \boldsymbol{\Psi}, \boldsymbol{\Sigma})$ satisfy that $\|\boldsymbol{\gamma}_j\| + \|\boldsymbol{\psi}_j\| < C$ and $\sigma_j^2 \in (C^{-1}, C)$ for $j = 1, \ldots, p_n$, and the limit of $p_n^{-1}\boldsymbol{\Psi}_0^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Psi}_0$ as $n \to \infty$ exists and has finite and positive eigenvalues. Furthermore, the true values of $\psi_{jj}^{(0,1)}$ $(j = 1, \ldots, r_0)$ and $\psi_{jj}^{(k)}$ $(j = 1, \ldots, r_k; k = 1, \ldots, K)$ are bounded below by C^{-1}. In addition, for $k = 1, \ldots, K$, all eigenvalues of $\widetilde{\boldsymbol{\Psi}}_0^{(k)\mathrm{T}}\widetilde{\boldsymbol{\Psi}}_0^{(k)}$ are bounded below by C^{-1}, where $\widetilde{\boldsymbol{\Psi}}_0^{(k)} \equiv (\boldsymbol{\Psi}_0^{(0,k)}, \boldsymbol{\Psi}_0^{(k)})$.

(C6) Let $\boldsymbol{S}^{(O)}$ be an arbitrary subvector of $\boldsymbol{S}$, $\boldsymbol{\nu} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi})$, ν_j be the jth component of $\boldsymbol{\nu}$ $(j = 1, \ldots, q_n)$, and q_n be the dimension of $\boldsymbol{\nu}$. Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be classes of

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

functions defined as

$$\mathcal{H}_1 = \left\{ h : (Y, \boldsymbol{X}, \boldsymbol{S}, \boldsymbol{\theta}) \mapsto \frac{\partial^2}{\partial \nu_j \partial \nu_k} \log f(Y; \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{X} + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{S}, \boldsymbol{\xi}); j, k = 1, \ldots, q_n \right\},$$

$$\mathcal{H}_2 = \left\{ h : (Y, \boldsymbol{X}, \boldsymbol{S}, \boldsymbol{\theta}) \mapsto \left\{ \frac{\partial}{\partial \nu_j} \log f(Y; \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{X} + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{S}, \boldsymbol{\xi}) \right\}^{k_0} \prod_{h=1}^{p_n} S_h^{k_h}; j = 1, \ldots, q_n; \right.$$

$$\left. k_0 = 0, 1, \text{or } 2; (k_1, \ldots, k_{p_n}) \text{ are nonnegative integers}; \sum_{h=1}^{p_n} k_h \leq 4 \right\},$$

and $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2$. Let $f(\boldsymbol{S} \mid Y, \boldsymbol{X}, \boldsymbol{S}^{(O)}; \boldsymbol{\theta})$ be the conditional density function of $\boldsymbol{S}$ given $(Y, \boldsymbol{X}, \boldsymbol{S}^{(O)})$, $\dot{\boldsymbol{f}}(\boldsymbol{S} \mid Y, \boldsymbol{X}, \boldsymbol{S}^{(O)}; \boldsymbol{\theta}) = \partial f(\boldsymbol{S} \mid Y, \boldsymbol{X}, \boldsymbol{S}^{(O)}; \boldsymbol{\theta}) / \partial \boldsymbol{\theta}$, and $\boldsymbol{S}^{(M)}$ be the vector of components of $\boldsymbol{S}$ that are not in $\boldsymbol{S}^{(O)}$. For $\widetilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$ within a small neighborhood of $\boldsymbol{\theta}_0$ and for all $h \in \mathcal{H}$, each component of

$$\mathrm{E} \int h(Y, \boldsymbol{X}, \boldsymbol{S}^{(O)}, \widetilde{\boldsymbol{\theta}}) \dot{\boldsymbol{f}}(\boldsymbol{S} \mid Y, \boldsymbol{X}, \boldsymbol{S}^{(O)}; \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{S}^{(M)}$$

is bounded by C, and $\mathrm{E}\{h(Y, \boldsymbol{X}, \boldsymbol{S}, \boldsymbol{\theta}) \mid Y, \boldsymbol{X}, \boldsymbol{S}^{(O)}; \boldsymbol{\theta}\}$ is four-times differentiable with respect to $\boldsymbol{\theta}$, with each component of its derivatives up to the fourth order uniformly bounded by a function of $(Y, \boldsymbol{X}, \boldsymbol{S}^{(O)})$ that has a finite second moment.

(C7) The score statistics for the outcome model satisfy $\mathrm{E}|Z(\boldsymbol{\alpha}_0^{\mathrm{T}} \boldsymbol{X} + \boldsymbol{\beta}_0^{\mathrm{T}} \boldsymbol{S}, \boldsymbol{\xi}_0) S_j|^k \leq k! C^k$, for $j = p_{1n} + 1, \ldots, p_n$ and $k \geq 2$.

(C8) For some $\eta \in (1 - \tau, 1 - \kappa)$,

$$\sup_{\boldsymbol{\theta}} \|\boldsymbol{I}(\boldsymbol{\theta})\|_{2,\infty} = O(n^{\tau + \eta - 1}),$$

where $\|\boldsymbol{D}\|_{2,\infty} = \| \sup_{\|\boldsymbol{v}\|=1} \boldsymbol{D}\boldsymbol{v} \|_\infty$, the supremum is taken in a small neighborhood of $\boldsymbol{\theta}_0$, and τ is defined in condition (C3).

(C9) The tuning parameter λ_n satisfies $\lambda_n n^{3/2 - \kappa - \eta} \to \infty$, where η is chosen in condition (C8).

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

Remark 1. Condition (C1) pertains to regularity conditions on the outcome model, and guarantees that with complete data, a local maximizer of the log-likelihood function is consistent for $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi})$. Condition (C2) requires that a nonvanishing proportion of subjects have complete data. Condition (C3) requires that the initial estimators of the nonzero parameters do not tend to zero at a rate faster than $n^{-\rho}$, whereas the estimators of the zero parameters are n^{τ}-consistent. This condition implies that the signal strength of S_j $(j = 1, \ldots, p_{1n})$ is bounded below by $Cn^{-\rho}$, for some positive constant C. Condition (C5) pertains to regularity conditions for consistent estimation of the latent factor model, and condition (C6) pertains to regularity conditions on the conditional density function of the missing data, given the observed data. Condition (C7) pertains to high-order moments of the score statistics of the outcome model, and condition (C8) is a general and weaker version of the strong irrepresentable condition (Zhao and Yu, 2006); these two conditions are imposed to ensure consistent model selection. Conditions (C4) and (C9) jointly require that $n^{1/2+\rho} \ll \lambda_n^{-1} \ll n^{3/2-\kappa-\eta}$. They ensure that the penalty for $\boldsymbol{\beta}$ is strong enough to impose model sparsity, but weak enough to yield consistent estimation of the nonzero parameters.

Remark 2. If we set the marginal estimators described in Section 2 as the initial estimators, then condition (C3) pertains to the relationships between the factor loadings across different features. Note that the marginal estimators are the MLE under the full likelihood (that incorporates the distribution of the incomplete variables), and thus are $n^{1/2}$-consistent for the "true" marginal regression parameters. In light of Proposition 1 of Fan and Song (2010), we can show that a marginal regression parameter under a gen-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

eralized linear model of Y tends to zero at a certain rate if and only if the correlation between the corresponding feature and $\boldsymbol{S}_{\mathcal{S}}^{\mathrm{T}}\boldsymbol{\beta}_{0\mathcal{S}}$ tends to zero at the same rate. Therefore, condition (C3) holds if $|\sum_{j=1}^{p_{1n}}\beta_{0j}\boldsymbol{\psi}_{0j}^{\mathrm{T}}\boldsymbol{\psi}_{0k}+\beta_{0k}\sigma_{0k}^2|>Cn^{-\rho}$ for $k=1,\ldots,p_{1n}$ and some positive constant C, and $\sum_{j=1}^{p_{1n}}\beta_{0j}\boldsymbol{\psi}_{0j}^{\mathrm{T}}\boldsymbol{\psi}_{0k}=O(n^{-\tau})$ for $k=p_{1n}+1,\ldots,p_n$, where for $j=1,\ldots,p_n$, $\boldsymbol{\psi}_{0j}^{\mathrm{T}}$ is the jth row of $\boldsymbol{\Psi}_0$, and σ_{0j}^2 is the jth diagonal element of $\boldsymbol{\Sigma}_0$.

Let $\boldsymbol{H}$ be the projection matrix onto the linear space of $\boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{\Psi}_0$. Our main theoretical results are summarized in the following theorem, the proof of which is given in Appendix A.2.

Theorem 1. *Under conditions (C1)–(C9), a local maximizer of $p\ell_n(\boldsymbol{\theta})$, denoted by $\widehat{\boldsymbol{\theta}}\equiv(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}}_{\mathcal{S}},\widehat{\boldsymbol{\beta}}_{\mathcal{N}},\widehat{\boldsymbol{\xi}},\widehat{\boldsymbol{\Gamma}},\widehat{\boldsymbol{\Psi}},\widehat{\boldsymbol{\Sigma}})$, satisfies that*

1. $\|\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0\|+\|\widehat{\boldsymbol{\beta}}_{\mathcal{S}}-\boldsymbol{\beta}_{0\mathcal{S}}\|+\|\widehat{\boldsymbol{\xi}}-\boldsymbol{\xi}_0\|+\|(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\Sigma}_0^{-1/2}(\widehat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma}_0)\|+\|(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\Sigma}_0^{-1/2}(\widehat{\boldsymbol{\Psi}}-\boldsymbol{\Psi}_0)\|+\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}_0\|=O_p(n^{-1/2}p_n^{1/2})$;
2. $\|\boldsymbol{H}\boldsymbol{\Sigma}_0^{-1/2}(\widehat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma}_0)\|+\|\boldsymbol{H}\boldsymbol{\Sigma}_0^{-1/2}(\widehat{\boldsymbol{\Psi}}-\boldsymbol{\Psi}_0)\|=O_p(n^{-1/2}p_n)$; *and*
3. $P(\widehat{\boldsymbol{\beta}}_{\mathcal{N}}=\boldsymbol{0})\to 1$ *as* $n\to\infty$.

Remark 3. This theorem provides the rate of convergence for the estimators of the nonzero parameters, and states that the estimators of the zero parameters are equal to zero with probability tending to one. A major step in the proof of Theorem 1 is to construct a shrinking neighborhood $\mathcal{N}$ of the true parameter values, and to show that $p\ell_n(\boldsymbol{\theta}_0)>\sup_{\boldsymbol{\theta}\in\partial\mathcal{N}}p\ell_n(\boldsymbol{\theta})$ with probability tending to one, where $\partial\mathcal{N}$ denotes the boundary of $\mathcal{N}$; similar approaches were adopted by Fan and Li (2001) and Fan and Peng (2004) to prove the consistency of the smoothly clipped absolute deviation estimator. The

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

proof is substantially more difficult for the factor model than for conventional regression models, because the largest eigenvalues of $(\boldsymbol{\Psi}\boldsymbol{\Psi}^{\mathrm{T}} + \boldsymbol{\Sigma})$ diverge to infinity (Bai and Liao, 2016). A key innovation in our proof is to identify the few eigenvalues of the Hessian matrix of $p\ell_n(\cdot)$ that tend to zero (as a result of the unboundedness of the eigenvalues of $(\boldsymbol{\Psi}\boldsymbol{\Psi}^{\mathrm{T}} + \boldsymbol{\Sigma})$), and to construct an "elliptical" $\mathcal{N}$ with diameter of order $n^{-1/2}p_n$ in directions that correspond to their eigenvectors, and with diameter of order $n^{-1/2}p_n^{1/2}$ in other directions. This construction guarantees that the second-order term in the Taylor series expansion of $p\ell_n(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_0$ is negative and dominates the first-order term at any $\boldsymbol{\theta} \in \partial\mathcal{N}$, with probability tending to one. Using this construction, we prove that the projections of $\boldsymbol{\Sigma}_0^{-1/2}\widehat{\boldsymbol{\Gamma}}$ and $\boldsymbol{\Sigma}_0^{-1/2}\widehat{\boldsymbol{\Psi}}$ along the linear space of $\boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{\Psi}_0$ are consistent at the $(n^{1/2}p_n^{-1})$-rate, whereas the estimators of all other nonzero parameters, including the regression parameters of interest, are consistent at the conventional $(n^{1/2}p_n^{-1/2})$-rate.

Remark 4. Bai and Li (2012) proved that the MLE of an unstructured factor loading matrix $\boldsymbol{\Psi}$ with complete data is $n^{-1/2}p_n^{1/2}$-consistent; this rate is faster than that given in Theorem 1. However, the arguments of Bai and Li (2012) are not applicable to our setting, especially when the missing-data mechanism depends on $\boldsymbol{S}$ (through Y). The asymptotic properties of the MLE of the factor model in the presence of missing data have not been previously studied, and it is unclear whether the convergence rates given in Theorem 1 can be improved.

Remark 5. The dimension of the features, p_n, that we allow is smaller than that in existing works on penalized regression with complete data (Fan and Peng, 2004; Fan and Lv, 2011), for three reasons. First, because the likelihood involves the model of the

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

incomplete variables, the number of (nonzero) parameters is much larger than that of existing works under the same number of features. Second, the rate of convergence of the proposed estimators is slower than that of existing works, such that a smaller p_n is required for the consistency of certain functions of the parameters, such as the information matrix. Third, owing to the structure of the factor model, the sizes of the derivatives of the log-likelihood function are potentially larger than those in conventional regression models, such that a smaller p_n is required to guarantee the concavity of the (observed) log-likelihood function around the true parameter values.

5. Simulation studies

We considered two types of features, $\boldsymbol{S}^{(1)}$ and $\boldsymbol{S}^{(2)}$, both with dimension $p/2$. We generated the features from the following factor model:

$$\boldsymbol{S}^{(k)} = \boldsymbol{\psi}^{(0,k)} U^{(0)} + \boldsymbol{\psi}^{(k)} U^{(k)} + \boldsymbol{\epsilon}^{(k)} \quad \text{for } k = 1, 2,$$

where $U^{(0)}$, $U^{(1)}$, and $U^{(2)}$ are independent standard normal variables, $\boldsymbol{\epsilon}^{(1)}$ and $\boldsymbol{\epsilon}^{(2)}$ are independent $(p/2)$-variate standard normal variables, and $\boldsymbol{\psi}^{(0,1)}$, $\boldsymbol{\psi}^{(0,2)}$, $\boldsymbol{\psi}^{(1)}$, and $\boldsymbol{\psi}^{(2)}$ are $(p/2)$-vectors of factor loadings. We set

$$\boldsymbol{\psi}^{(0,1)} = \boldsymbol{\psi}^{(0,2)} = (\underbrace{0.2, \ldots, 0.2}_{20 \text{ terms}}, \underbrace{-0.2, \ldots, -0.2}_{(p/4-10) \text{ terms}}, \underbrace{0.2, \ldots, 0.2}_{(p/4-10) \text{ terms}})^{\mathrm{T}},$$

$$\boldsymbol{\psi}^{(1)} = \boldsymbol{\psi}^{(2)} = (\underbrace{0.4, \ldots, 0.4}_{20 \text{ terms}}, \underbrace{0.4, \ldots, 0.4}_{(p/4-10) \text{ terms}}, \underbrace{-0.2, \ldots, -0.2}_{(p/4-10) \text{ terms}})^{\mathrm{T}}.$$

In this setting, the first 20 components of each type of features are positively associated with both the common and the type-specific latent variables, whereas the remaining components are negatively associated with either the common or the type-specific latent

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

variable. As a result, the first 20 components of each type of features are relatively strongly associated with each other, but they are weakly associated with other features. We let the outcome variable Y be continuous or binary. For the continuous case, we set $Y = \sum_{j=1}^{15} 0.05(S_j^{(1)} + S_j^{(2)}) + \delta$, where δ follows a standard normal distribution, and $S_j^{(k)}$ is the jth component of $\boldsymbol{S}^{(k)}$ $(k = 1, 2)$. For the binary case, we set $P(Y = 1 \mid \boldsymbol{S}^{(1)}, \boldsymbol{S}^{(2)}) = \text{logit}^{-1}\{-3 + \sum_{j=1}^{15} 0.15(S_j^{(1)} + S_j^{(2)})\}$, such that $P(Y = 1) \approx 0.1$. We set $\boldsymbol{S}^{(1)}$ to be completely observed and $\boldsymbol{S}^{(2)}$ to be missing for 50% of the subjects, based on one of the following missing-data mechanisms: (1) missing completely at random (MCAR); and (2) missing at random (MAR), such that in the case of the continuous outcome variable, $\boldsymbol{S}^{(2)}$ is observed for subjects with extreme values of Y, and in the case of the binary outcome variable, $\boldsymbol{S}^{(2)}$ is observed for all subjects with $Y = 1$ and for a random subset of subjects with $Y = 0$.

We adopted two penalization methods: the lasso (Tibshirani, 1996) and adaptive lasso (Zou, 2006). For the lasso, we set each weight term w_j to one. For the adaptive lasso, we used the MLE of the regression parameters in the marginal regression models of Y against $S_j^{(k)}$ $(j = 1, \ldots, p/2; k = 1, 2)$ as the initial estimators. The tuning parameter for each method was selected using five-fold cross-validation, where the cross-validation error is defined as the negative log-likelihood value, and the grid for λ is $\{(0.01)^{j/100}\lambda_{\max}\}_{j=0,1,\ldots,99}$. For each penalization method, we considered four methods for handling missing data: (1) variable selection on complete cases only; (2) variable selection with missing data (singly) imputed using the structured matrix completion method of Cai, Cai, and Zhang (2016), where the row thresholding parameter is set to $2\{p/\text{length}(\boldsymbol{S}^{(2)})\}^{1/2} \approx 2.828$; (3) variable selection with missing data

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

imputed by the posterior expectation under the proposed factor model, with the factor model estimated using only $(\boldsymbol{S}^{(1)}, \boldsymbol{S}^{(2)})$; and (4) the proposed penalized-likelihood method. For method (4) and the estimation of the factor model in method (3), we used the function GQN2_ORDER of the C++ software SPARSE_GRID_HW (available at https://people.sc.fsu.edu/~jburkardt/cpp_src/sparse_grid_hw/sparse_grid_hw.html) and set level = 3 to generate the nodes and weights for the sparse-grid quadrature. We terminated the EM algorithm when the maximum absolute difference between the parameter estimators of two consecutive iterations became smaller than 10^{-5}. For methods (3) and (4), we considered models with $(r_0, r_1, r_2) \in \{(r_0, r_1, r_2) : \sum_{k=0}^{2} |r_k - 1| \leq 1, r_0 \neq 0\}$, and selected the model using the Bayesian information criterion (BIC) (Schwarz, 1978). Note that the models considered differ from the true model by at most one latent variable. Models with $r_0 = 0$ assume independence between the two types of features, and thus were not considered. In all models, we set $X = 1$.

We set $n = 500$ and $p = 100$ or 300. For each method, we report the number of variables selected, false discovery rate, true positive rate, and prediction error. The false discovery rate is defined as the proportion of selected variables that have a zero true parameter value, and the true positive rate is defined as the proportion of variables with nonzero true parameter values that are selected. The prediction error is defined as $\mathrm{E}\{\boldsymbol{S}^{(1)\mathrm{T}}(\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}_0^{(1)}) + \boldsymbol{S}^{(2)\mathrm{T}}(\widehat{\boldsymbol{\beta}}^{(2)} - \boldsymbol{\beta}_0^{(2)})\}^2$, where $\widehat{\boldsymbol{\beta}}^{(k)}$ and $\boldsymbol{\beta}_0^{(k)}$ are the estimated and the true values, respectively, of the regression parameters of $\boldsymbol{S}^{(k)}$ $(k = 1, 2)$. For the single imputation method based on the factor model and the proposed method, we also report the proportion of replicates in which the correct numbers of latent variables are selected. The results, which are based on 200 replicates, are summarized in Tables 1 and 2.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

Table 1: Simulation results for the continuous outcome variable

	Lasso					A-Lasso				
	Variables selected	FDR	TPR	Pred error	Correct model	Variables selected	FDR	TPR	Pred error	Correct model
MCAR; $p = 100$										
Complete	26.4	0.372	0.531	0.125	N/A	22.4	0.280	0.526	0.122	N/A
SMC	28.9	0.377	0.584	0.120	N/A	23.7	0.276	0.555	0.122	N/A
Imputed	30.2	0.354	0.632	0.098	1	25.0	0.243	0.616	0.096	1
Proposed	25.6	0.294	0.592	0.105	1	23.0	0.218	0.591	0.098	1
MAR; $p = 100$										
Complete	32.0	0.350	0.670	0.160	N/A	26.3	0.231	0.661	0.206	N/A
SMC	28.8	0.355	0.589	0.150	N/A	22.2	0.203	0.570	0.152	N/A
Imputed	31.7	0.371	0.644	0.134	1	25.3	0.239	0.629	0.149	1
Proposed	27.9	0.269	0.668	0.089	1	24.5	0.186	0.654	0.080	1
MCAR; $p = 300$										
Complete	29.9	0.497	0.454	0.156	N/A	35.6	0.533	0.507	0.168	N/A
SMC	34.3	0.485	0.547	0.130	N/A	32.4	0.428	0.577	0.123	N/A
Imputed	34.0	0.462	0.567	0.124	0.910	36.9	0.465	0.607	0.130	0.910
Proposed	24.2	0.366	0.495	0.140	0.910	27.6	0.367	0.556	0.120	0.915
MAR; $p = 300$										
Complete	36.9	0.473	0.600	0.153	N/A	34.2	0.412	0.626	0.236	N/A
SMC	40.1	0.504	0.616	0.157	N/A	35.9	0.408	0.649	0.198	N/A
Imputed	34.0	0.462	0.568	0.151	0.995	34.4	0.427	0.615	0.189	0.995
Proposed	26.5	0.320	0.586	0.118	0.995	27.2	0.294	0.615	0.094	0.995

NOTE: "A-Lasso" stands for adaptive lasso; "Complete," "SMC," "Imputed," and "Proposed" stand for the complete-case analysis, the structured matrix completion method of Cai, Cai, and Zhang (2016), single imputation based on the factor model, and the proposed method, respectively; "Variables selected," "FDR," "TPR," "Pred error," and "Correct model" stand for the average number of variables selected, the false discovery rate, the true positive rate, the prediction error, and the proportion of replicates in which the correct factor model is selected, respectively.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

Table 2: Simulation results for the binary outcome variable

	Lasso					A-Lasso				
	Variables selected	FDR	TPR	Pred error	Correct model	Variables selected	FDR	TPR	Pred error	Correct model
MCAR; $p = 100$										
Complete	20.0	0.356	0.408	1.486	N/A	17.1	0.286	0.396	1.507	N/A
SMC	23.2	0.364	0.468	1.451	N/A	18.8	0.274	0.439	1.370	N/A
Imputed	24.8	0.343	0.522	1.214	1	20.7	0.251	0.504	1.113	1
Proposed	24.4	0.353	0.506	1.277	1	20.5	0.251	0.502	1.100	1
MAR; $p = 100$										
Complete	25.8	0.340	0.550	1.134	N/A	20.4	0.227	0.515	1.075	N/A
SMC	23.7	0.338	0.498	1.464	N/A	19.0	0.220	0.481	1.530	N/A
Imputed	25.0	0.337	0.534	1.272	1	20.5	0.232	0.513	1.430	1
Proposed	27.2	0.341	0.577	1.101	1	21.8	0.227	0.552	0.966	1
MCAR; $p = 300$										
Complete	23.4	0.508	0.338	1.827	N/A	26.8	0.535	0.388	2.300	N/A
SMC	28.4	0.495	0.432	1.592	N/A	27.6	0.462	0.460	1.462	N/A
Imputed	29.1	0.488	0.452	1.515	0.915	30.4	0.478	0.490	1.488	0.915
Proposed	27.5	0.487	0.426	1.624	0.880	30.7	0.486	0.484	1.389	0.910
MAR; $p = 300$										
Complete	30.9	0.498	0.484	1.371	N/A	27.4	0.422	0.499	1.379	N/A
SMC	31.4	0.488	0.494	1.502	N/A	28.1	0.409	0.523	1.975	N/A
Imputed	25.7	0.440	0.446	1.531	0.990	27.3	0.425	0.491	2.003	0.990
Proposed	31.5	0.479	0.511	1.346	0.990	30.8	0.443	0.540	1.182	0.990

NOTE: See NOTE to Table 1.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

With the sparse-grid numerical integration, the mean computing time for a single E-step is about 0.1–0.2 seconds under $p = 100$, and about 0.3–0.7 seconds under $p = 300$, in various settings. In all scenarios, the proposed method performs substantially better than the complete-case analysis, because the latter discards subjects with partial information, and thus is less efficient. Single imputation based on the factor model has overall better variable selection and prediction performance than that based on structured matrix completion, because the former assumes a correct imputation model. Under MCAR, the proposed method and the single imputation method based on the factor model perform similarly, possibly because the structure of the factor model can be accurately recovered using $(\boldsymbol{S}^{(1)}, \boldsymbol{S}^{(2)})$ alone. Under MAR, however, the proposed method yields a substantially smaller prediction error and similar or better false discovery and true positive rates than those of the single imputation methods, owing to either the estimation bias of the factor model or the failure to recover the low-rank structure underlying $(\boldsymbol{S}^{(1)}, \boldsymbol{S}^{(2)})$ in the single imputation methods. Note that, in general, the proposed method yields better results under MAR than under MCAR, because subjects with extreme outcome values contain more information than do randomly selected subjects. Owing to estimation bias, however, the single imputation methods fail to capture the extra information and perform worse under MAR. For the single imputation method based on the factor model and the proposed method, the BIC selects the correct factor model in the majority of replicates.

In general, the adaptive lasso yields a lower false discovery rate and a smaller prediction error than those of the lasso; in some cases, the lasso yields a higher true positive rate than that of the adaptive lasso, probably because the lasso selects more variables. These results agree with the expectation that by assigning larger penalties to less important

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

variables, the adaptive lasso outperforms the lasso.

We conducted additional simulation studies to investigate the performance of the proposed methods under a misspecified latent factor model. We show that when every feature depends on all three latent variables, the proposed method with $r_0 = r_1 = r_2 = 1$ still yields superior performance over the complete-case analysis and imputation methods. Details of the simulation settings and results are presented in Section S1 of the Supplementary Material.

6. A real study

We considered the TCGA data (available at http://gdac.broadinstitute.org/) on two smoking-related, upper aerodigestive tract cancers: head and neck squamous cell carcinoma (HNSC) and lung adenocarcinoma (LUAD). After removing patients with missing clinical data, the total sample size was 955, with 448 HNSC patients and 507 LUAD patients. We considered the outcome variable tumor stage, dichotomized into stage I/II and stage III/IV. The proportions of patients with later stages were 0.77 for HNSC and 0.22 for LUAD. We considered two types of genomic features, namely, gene expressions and protein expressions, with 18028 and 155 variables, respectively. A total of 400 patients had no protein data, owing to insufficient tissue sample left for protein expression measurement, so the missing mechanism does not depend on the protein expression values. In addition, patients missed 763 gene expressions, on average, and among patients with some protein data, an average of 4.3 protein expressions are missing; the missing data did not exhibit a blockwise pattern. The missing-data pattern is plotted in Figure S1 in the Supplementary Material.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

We first screened the gene expressions according to their marginal associations with tumor stage, such that the resulting number of variables and sample size were comparable. We tested each gene expression's marginal association with tumor stage using the score test (adjusted for cancer type), with the model for missing data included in the likelihood, and selected the 500 variables with the smallest p-values. Then, we fit a logistic regression model for tumor stage, with cancer type as X and the screened gene expressions and protein expressions as $\boldsymbol{S}$. For the latent factor model, we set cancer type as a covariate and ranged the total number of latent variables r from three to six with $(r_0, r_1, r_2) \in \{(r_0, r_1, r_2) : \sum_{k=0}^{2} r_k \leq 6, r_k \geq 1 \text{ for } k = 0, 1, 2\}$. We used the adaptive-lasso penalty with the marginal regression parameter estimators as the initial estimators and selected the tuning parameter using five-fold cross-validation.

The BIC picked $(r_0, r_1, r_2) = (2, 3, 1)$ for the factor model. A total of 53 genomic features were selected, with 45 gene expressions and eight protein expressions; the selected features are presented in Table S3 in the Supplementary Material. Several selected genes, including WDR37, FUT7, and DDIT4, were previously reported to be associated with metastasis or patient survival (Ogawa, Inoue, and Koide, 1997; Läubli et al., 2006; Wang et al., 2015). The selected proteins include the epidermal growth factor receptor, which is known to be involved in the pathogenesis and progression of different types of cancer (Nicholson, Gee, and Harper, 2001; Normanno et al., 2006).

The estimated outcome model and factor model enabled us to construct a personal genomic risk score. For each patient, we calculated the posterior expectation of the latent variable, denoted by $\widehat{\boldsymbol{U}}$, and imputed the missing values of the genomic features by the corresponding element of $\boldsymbol{\Gamma X} + \boldsymbol{\Psi}\widehat{\boldsymbol{U}}$, where the parameters were evaluated at their

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

estimated values. The risk score is defined as $\boldsymbol{\beta}^{\mathrm{T}}\widehat{\boldsymbol{S}}$, with $\boldsymbol{\beta}$ evaluated at its estimated value and $\widehat{\boldsymbol{S}}$ the vector of the observed or imputed genomic features.

We evaluated the association between the risk score and progression-free survival time (since initial diagnosis). For each cancer type, we fit a stratified Cox model of progression-free survival time against the risk score, stratified by tumor stage. The likelihood-ratio p-values for the effects of the risk score are 4.20×10^{-4} and 6.41×10^{-6} for HNSC and LUAD, respectively. These p-values are highly significant, suggesting that the selected genomic features are highly relevant to cancer progression. Note that the results are not due to overfitting, because progression-free survival time is not involved in the calculation of the risk score, and all evaluations are stratified by tumor stage and cancer type.

For comparison, we performed similar analyses using imputed data and the complete cases only. (The method developed by Cai, Cai, and Zhang (2016) is not applicable to the general missing-data pattern exhibited in this data set.) For the single imputation, we imputed the missing values using k-nearest-neighbor imputation (Troyanskaya et al., 2001) with $k = 10$, and performed the adaptive lasso. For the complete-case analysis, because very few patients have complete data, we removed only those patients with no protein expression data; we imputed the missing values for the remaining patients using 10-nearest-neighbor imputation, and performed the adaptive lasso on the imputed data set. In constructing the risk scores, we again imputed missing values of $\boldsymbol{S}$ using 10-nearest-neighbor imputation.

For the complete-case analysis, the likelihood-ratio p-values for the association between the risk score and progression-free survival time under the stratified Cox model are 0.024 and 3.98×10^{-6} for HNSC and LUAD, respectively. While the p-value for LUAD is as

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

significant as that of the proposed method, the p-value for HNSC is only mildly significant. For the single imputation, the likelihood-ratio p-values are 1.84×10^{-3} and 1.80×10^{-5} for HNSC and LUAD, respectively. Single imputation yields similar results to those of the proposed method, with slightly less significant p-values for both cancer types.

As a by-product of the proposed method, we obtained a low-dimensional projection of the genomic features, $\widehat{\boldsymbol{U}}$. We constructed an alternative risk score, defined as $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Psi}\widehat{\boldsymbol{U}}$, which is the estimated effect of the projected genomic features on tumor stage. For each cancer type and tumor stage group, we divided patients into two equal-sized risk groups according to their risk scores. The Kaplan–Meier curves of the progression-free survival times for the risk groups are given in Figure 1. We also tested the association between the risk score and progression-free survival time under the stratified Cox model, and the likelihood-ratio p-values are 1.55×10^{-5} and 2.22×10^{-3} for HNSC and LUAD, respectively. Remarkably, patients classified into high-risk groups tend to have lower (progression-free) survival probabilities than patients in the corresponding low-risk groups. In addition, the likelihood-ratio tests are significant for both cancer types, and the p-value for HNSC is smaller than those obtained from the original risk score (using the proposed method, single imputation, or complete-case analysis). A possible explanation is that the projection of the genomic features contains less noise than do the individual genomic features, and thus better represents patients' genomic characteristics.

Finally, we compared the three methods for handling missing data using cross-validation. We first independently sampled 10 training sets from the full data set. Each training set consists of 60% of the patients ($n = 573$); the training set and the full data set have approximately equal distributions of cancer type, tumor stage, and missing-data propor-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

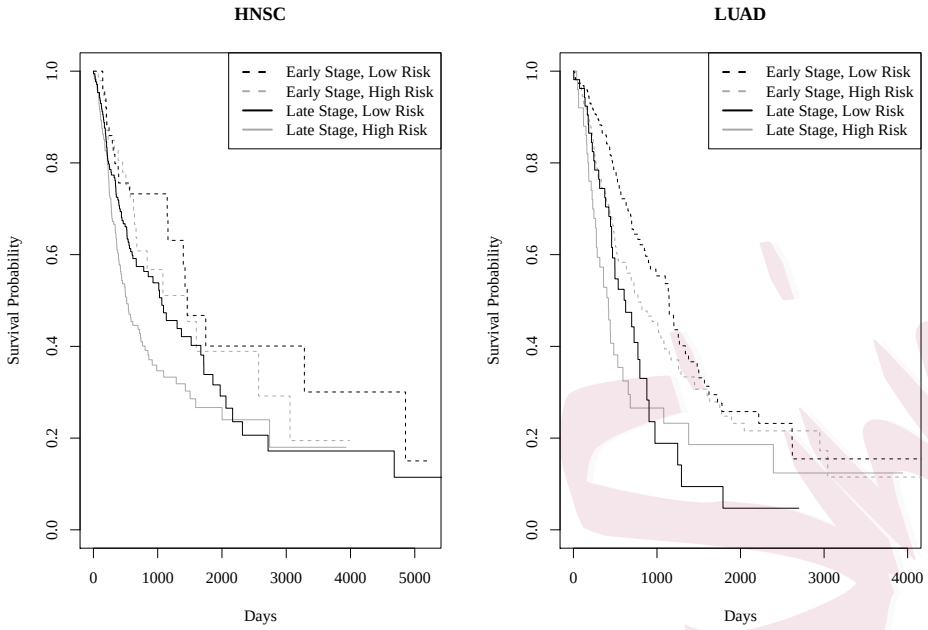

Figure 1: Kaplan–Meier curves for risk groups defined by the projected genomic features.

tion. In each training set, we followed the procedures used in the analyses of the whole data set: we screened the gene expressions by selecting the 500 gene expressions with the strongest association with tumor stage (after adjusting for cancer type), and performed a complete-case analysis, single imputation by 10-nearest-neighbor imputation, and the proposed method on the screened variables over a sequence of tuning parameter values. For the proposed method, we set the numbers of latent variables in the factor model to $r_0 = 2$, $r_1 = 3$, and $r_2 = 1$.

To evaluate the performance of each method, we used 10-nearest-neighbor imputation to impute the missing values in the full data set, and let each validation set consist of patients that were not part of a corresponding training set. For each training and validation data split and each method, we performed the analysis using the training

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

set over a series of tuning parameter values, and computed the area under the receiver operating characteristic curve (AUC) of the resulting risk scores in the validation set. Each method produces a curve of AUC against tuning parameter values for each data split. Because the variables in the regression analyses differ across data splits owing to screening, it is not appropriate to combine the results over splits at the same tuning parameter values. Instead, we averaged the AUC values over splits at models with the same number of selected variables.

The complete-case analysis yields the smallest average AUC at any given model size, and the average AUC values between the single imputation and the proposed method are similar. Specifically, the maximum average AUC values for the complete-case analysis, single imputation, and the proposed method are 0.802, 0.816, and 0.817, respectively. Possibly because of the small sample size of the training data sets, the single imputation and the proposed method perform similarly.

7. Discussion

We have proposed a penalized-likelihood approach to variable selection and parameter estimation for multiple types of many features with missing data. Our approach accommodates arbitrary missing-data patterns, including, but not restricted to blockwise missing data. We prove the estimation and model selection consistency of the penalized estimator, and develop an efficient EM algorithm for its computation. A key advantage of the proposed estimator is that it is consistent under MAR, whereas single imputation is biased, in general, when the missing-data mechanism depends on the outcome variable.

The structure of the latent factor model facilitates efficient computation of the penal-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

ized estimator under general missing-data patterns. Under the factor model, the conditional distribution of $\boldsymbol{S}^{(M)}$ given $(\boldsymbol{U}, \boldsymbol{S}^{(O)})$ is equal to the conditional distribution of $\boldsymbol{S}^{(M)}$ given $\boldsymbol{U}$ alone. This structure makes it simpler to evaluate the conditional distribution of the variables with missing values given the observed variables in the E-step. (By contrast, if an unstructured covariance matrix for $\boldsymbol{S}$ is assumed, then the conditional expectation may involve inversions of large matrices.) In addition, because of the conditional independence of components of $\boldsymbol{S}$, the factor loadings for each component of $\boldsymbol{S}$ can be updated separately with closed-form solutions in the M-step.

Under the normality assumption on the error term $\boldsymbol{\epsilon}$ in the factor model, the EM algorithm involves only a low-dimensional numerical integration, even when the dimension of the incomplete variables is high. This assumption ensures the existence of a linear transformation of $\boldsymbol{\epsilon}$, denoted by $\widetilde{\boldsymbol{\epsilon}}$, such that the components of $\widetilde{\boldsymbol{\epsilon}}$ are independent, and the outcome variable depends on $\boldsymbol{\epsilon}$ only through a single component of $\widetilde{\boldsymbol{\epsilon}}$. For general distributions of $\boldsymbol{\epsilon}$, such a linear transformation is not available, and the integration in the likelihood function cannot be simplified.

We use the BIC to select the numbers of common and type-specific latent variables. When the number of feature types is large, this approach may involve evaluating a large number of models, and may be computationally intensive. Alternatively, we can consider penalization approaches similar to those of Ibrahim et al. (2011) and Caner and Han (2014), which penalize the variances of the latent variables or columns of the factor loading matrix.

One application of the proposed approach lies in the dimension reduction of multimodality features. In genomics studies, low-dimensional projections of genomic features

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

have been used to pick out technical errors (Leek et al., 2010), characterize the activities of gene sets (Fan et al., 2011), and discover molecular subtypes of patients (Shen, Olshen, and Ladanyi, 2009; Shen, Wang, and Mo, 2013). Most projection methods do not account for missing data, much less a missing-not-at-random mechanism. Under the proposed framework, if the missing-data mechanism depends on an external variable (that is associated with $\boldsymbol{S}$), then we can set the variable as the outcome variable Y, and the resulting factor model can be estimated without bias and be used to generate the projection.

Our work can be extended in several directions. First, we may consider a potentially right-censored outcome variable that follows the Cox proportional hazards model. This extension would find applications in many multi-platform genomics studies in which the outcome of interest is time to death or disease progression. For this extension, the penalized estimator can be computed using the EM algorithm, where the M-step maximizes a quadratic approximation of the (expected) log-partial likelihood using the coordinate-descent algorithm (Simon et al., 2011).

Another extension is to allow associations between features beyond those explained by the latent variables. The extra associations can be accommodated by adopting an approximate factor model of $\boldsymbol{S}$, which allows nonzero, but sparse off-diagonal elements in $\boldsymbol{\Sigma}$. A sparse estimator of $\boldsymbol{\Sigma}$ can be obtained by including a penalty term on the off-diagonal elements of the covariance matrix in the penalized likelihood (Bai and Liao, 2016). This generalization, however, imposes considerable computational challenges by complicating the conditional distribution of $\boldsymbol{S}$ given $\boldsymbol{U}$, and introducing an extra tuning parameter.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

A third extension is to consider nonnormal incomplete features. We may fit a semiparametric factor model, such that each latent variable is a nonparametric monotone transformation of a Gaussian variable. We can adopt penalized sieve maximum likelihood methods for estimation and the EM algorithm for computation. Another possibility is to assume that transformations of components of $\boldsymbol{S}$ follow the (Gaussian) factor model, and then fit a regression model of Y on the transformed $\boldsymbol{S}$.

Finally, we have only established the estimation and selection consistency of the penalized estimator. It is plausible that the estimator for the nonzero parameters exhibits the so-called oracle properties (Fan and Li, 2001) under further regularity conditions. However, inference based on the oracle properties ignores the variability arising from variable selection, and thus is seldom conducted in practice. A possible future research direction is to investigate a "de-biased" version of the penalized estimator (van de Geer et al., 2014; Zhang and Zhang, 2014), which yields more reliable inference at the expense of nonsparsity of the estimators.

Supplementary Material

The Supplementary Material contains additional theoretical results, simulation results, and details of the real-data analysis.

Acknowledgments

This research was supported by a start-up research grant from the Hong Kong Polytechnic University (1-BE02), Hong Kong Research Grants Council grant PolyU 253042/18P, and National Institutes of Health grants R01HG009974, R01GM047845, and P01CA142538.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

Appendix A: Computational and technical details

A.1 Details of the EM algorithm

To express the first and second conditional moments of $\widetilde{\boldsymbol{S}}_i$ given $\mathcal{O}_i \equiv (Y_i, \boldsymbol{X}_i, \boldsymbol{S}_i^{(O)})$ in terms of those of $(\boldsymbol{U}_i, \widetilde{\epsilon}_i)$, we define $\widetilde{\boldsymbol{\beta}}_i^{(M)} = (\boldsymbol{\beta}_i^{(M)})_{j:\beta_{ij}^{(M)}\neq 0}$, $\widetilde{\boldsymbol{\sigma}}_i^{(M)} = (\boldsymbol{\sigma}_i^{(M)})_{j:\beta_{ij}^{(M)}\neq 0}$, and $\widetilde{p}_i^{(M)}$ to be the dimension of $\widetilde{\boldsymbol{\beta}}_i^{(M)}$, where $\boldsymbol{\sigma}_i^{(M)} = (\sigma_1, \ldots, \sigma_{p_n})^{\mathrm{T}}_{\mathcal{M}_i}$. Let $\widetilde{\boldsymbol{\vartheta}}_{i1} = c_i^{-1}(\widetilde{\beta}_{i1}^{(M)}\widetilde{\sigma}_{i1}^{(M)}, \ldots, \widetilde{\beta}_{i\widetilde{p}^{(M)}}^{(M)}\widetilde{\sigma}_{i\widetilde{p}^{(M)}}^{(M)})^{\mathrm{T}}$, and set $(\widetilde{\boldsymbol{\vartheta}}_{i2}, \ldots, \widetilde{\boldsymbol{\vartheta}}_{i\widetilde{p}_i^{(M)}})$ to be $\widetilde{p}_i^{(M)}$-dimensional unit vectors that are orthogonal to $\widetilde{\boldsymbol{\vartheta}}_{i1}$ and to each other. The first and second (conditional) moments of $\widetilde{\boldsymbol{S}}_i$ involve $\mathrm{E}(\boldsymbol{U}_i \mid \mathcal{O}_i)$, $\mathrm{E}(\boldsymbol{\epsilon}_i^{(M)} \mid \mathcal{O}_i)$, $\mathrm{E}(\boldsymbol{U}_i\boldsymbol{U}_i^{\mathrm{T}} \mid \mathcal{O}_i)$, $\mathrm{E}(\widehat{\boldsymbol{\Psi}}_i^{(M)}\boldsymbol{U}_i\boldsymbol{\epsilon}_i^{(M)\mathrm{T}} \mid \mathcal{O}_i)$, and $\mathrm{E}(\boldsymbol{\epsilon}_i^{(M)}\boldsymbol{\epsilon}_i^{(M)\mathrm{T}} \mid \mathcal{O}_i)$. The moments of $\boldsymbol{U}_i$ are readily available from (3.2) by setting $g(\boldsymbol{U}_i, \widetilde{\epsilon}_i) = \boldsymbol{U}_i$ or $\boldsymbol{U}_i\boldsymbol{U}_i^{\mathrm{T}}$. For the other terms in the moments of $\widetilde{\boldsymbol{S}}_i$, we have

$$\mathrm{E}(\epsilon_{ij} \mid \mathcal{O}_i) = \begin{cases} c_i^{-1}\sigma_{ij}^{(M)}\widetilde{\vartheta}_{i1,m_i(j)}\mathrm{E}(\widetilde{\epsilon}_i \mid \mathcal{O}_i) & \text{if } \beta_{ij}^{(M)} \neq 0, \\ 0 & \text{otherwise,} \end{cases}$$

$$\mathrm{E}(\widehat{\boldsymbol{\Psi}}_i^{(M)}\boldsymbol{U}_i\boldsymbol{\epsilon}_i^{(M)\mathrm{T}} \mid \mathcal{O}_i)_{jk} = \begin{cases} c_i^{-1}\sigma_{ik}^{(M)}\widetilde{\vartheta}_{i1,m_i(k)}\sum_{h=1}^{r}\widehat{\psi}_{ijh}^{(M)}\mathrm{E}(U_{ih}\widetilde{\epsilon}_i \mid \mathcal{O}_i) & \text{if } \beta_{ik}^{(M)} \neq 0, \\ 0 & \text{otherwise,} \end{cases}$$

$$\mathrm{E}(\boldsymbol{\epsilon}_i^{(M)}\boldsymbol{\epsilon}_i^{(M)\mathrm{T}} \mid \mathcal{O}_i)_{jk} = \begin{cases} c_i^{-2}\sigma_{ij}^{(M)}\sigma_{ik}^{(M)}\delta_{jk} & \text{if } \beta_{ij}^{(M)} = 0 \text{ or } \beta_{ik}^{(M)} = 0, \\ c_i^{-2}\sum_{h=1}^{\widetilde{p}_i^{(M)}}\{I(h=1)\mathrm{E}(\widetilde{\epsilon}_i^2 \mid \mathcal{O}_i) & \\ +I(h\neq 1)\}\sigma_{ij}^{(M)}\sigma_{ik}^{(M)}\widetilde{\vartheta}_{ih,m_i(j)}\widetilde{\vartheta}_{ih,m_i(k)} & \text{otherwise,} \end{cases}$$

where $\widetilde{\vartheta}_{ijk}$ is the kth component of $\widetilde{\boldsymbol{\vartheta}}_{ij}$, $m_i(j)$ is such that $\widetilde{\beta}_{i,m_i(j)}^{(M)} = \beta_{ij}^{(M)}$, $\widehat{\psi}_{ijk}^{(M)}$ is the (j,k)th element of $\widehat{\boldsymbol{\Psi}}_i^{(M)}$, and $\delta_{jk} = I(j=k)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

A.2 Proofs of technical results

Before proving Theorem 1, we present the following lemma, which pertains to an estimator of the nonzero parameters. Let $\widehat{\boldsymbol{\theta}}_{\text{Oracle}} \equiv (\widehat{\boldsymbol{\alpha}}_{\text{Oracle}}, \widehat{\boldsymbol{\beta}}_{\text{Oracle}}, \widehat{\boldsymbol{\xi}}_{\text{Oracle}}, \widehat{\boldsymbol{\Gamma}}_{\text{Oracle}}, \widehat{\boldsymbol{\Psi}}_{\text{Oracle}}, \widehat{\boldsymbol{\Sigma}}_{\text{Oracle}})$ be a local maximizer of $p\ell_n(\boldsymbol{\theta})$ when $\boldsymbol{\beta}_{\mathcal{N}}$ is fixed at $\mathbf{0}$. For any potentially random real-valued sequences a_n and b_n, we say that a_n is dominated by b_n if $|a_n/b_n| = o_p(1)$. We have the following result about the oracle estimator.

Lemma 1. *Under conditions (C1)–(C6), there is a version of $\widehat{\boldsymbol{\theta}}_{\text{Oracle}}$ that satisfies*

$$\begin{aligned}
&\|\widehat{\boldsymbol{\alpha}}_{\text{Oracle}} - \boldsymbol{\alpha}_0\| + \|\widehat{\boldsymbol{\beta}}_{\text{Oracle}} - \boldsymbol{\beta}_0\| + \|\widehat{\boldsymbol{\xi}}_{\text{Oracle}} - \boldsymbol{\xi}_0\| + p_n^{-1/2}\|\boldsymbol{H}\boldsymbol{\Sigma}_0^{-1/2}(\widehat{\boldsymbol{\Gamma}}_{\text{Oracle}} - \boldsymbol{\Gamma}_0)\| \\
&+ \|(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\Sigma}_0^{-1/2}(\widehat{\boldsymbol{\Gamma}}_{\text{Oracle}} - \boldsymbol{\Gamma}_0)\| + p_n^{-1/2}\|\boldsymbol{H}\boldsymbol{\Sigma}_0^{-1/2}(\widehat{\boldsymbol{\Psi}}_{\text{Oracle}} - \boldsymbol{\Psi}_0)\| \\
&+ \|(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\Sigma}_0^{-1/2}(\widehat{\boldsymbol{\Psi}}_{\text{Oracle}} - \boldsymbol{\Psi}_0)\| + \|\widehat{\boldsymbol{\Sigma}}_{\text{Oracle}} - \boldsymbol{\Sigma}_0\| = O_p(n^{-1/2}p_n^{1/2}).
\end{aligned}$$

The proof of Lemma 1 is presented in Section S3 of the supplementary materials.

Proof of Theorem 1. By Lemma 1, $\widehat{\boldsymbol{\theta}}_{\text{Oracle}}$ converges in probability to $\boldsymbol{\theta}_0$ at the desired rate of convergence, so it suffices to show that there exists a local maximizer of $p\ell_n(\boldsymbol{\theta})$, denoted by $\widehat{\boldsymbol{\theta}}$, such that $P(\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_{\text{Oracle}}) \to 1$ for some version of the oracle estimator $\widehat{\boldsymbol{\theta}}_{\text{Oracle}}$. Let $\boldsymbol{\theta}_{\mathcal{S}}$ be the vector that consists of $(\boldsymbol{\alpha}, \boldsymbol{\beta}_{\mathcal{S}}, \boldsymbol{\xi}, \boldsymbol{\Gamma}, \boldsymbol{\Psi}, \boldsymbol{\Sigma})$. The desired result follows if the following Karush–Kuhn–Tucker conditions hold:

$$\begin{aligned}
&\left\{\mathbb{P}_n \frac{\partial}{\partial \boldsymbol{\theta}_{\mathcal{S}}}\ell(\boldsymbol{\theta}) - \lambda_n \boldsymbol{w}_{\mathcal{S}} \circ \operatorname{sgn}(\boldsymbol{\beta}_{\mathcal{S}})\right\}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\text{Oracle}}} = \mathbf{0}, \\
&\left|\mathbb{P}_n \frac{\partial}{\partial \beta_j}\ell(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\text{Oracle}}}\right| < \lambda_n w_j \qquad \text{for } j = p_{1n}+1, \ldots, p_n, \\
&\lambda_{\min}\left\{-\mathbb{P}_n \frac{\partial^2}{\partial \boldsymbol{\theta}_{\mathcal{S}} \partial \boldsymbol{\theta}_{\mathcal{S}}^{\mathrm{T}}}\ell(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\text{Oracle}}}\right\} > 0,
\end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

where $\ell(\boldsymbol{\theta})$ denotes the log-likelihood function for a single subject, $\mathbb{P}_n$ denotes the empirical process measure, and $\mathrm{sgn}(\boldsymbol{\beta}_{\mathcal{S}})$ is the vector of the signs of $\boldsymbol{\beta}_{\mathcal{S}}$. The first condition holds by the definition of $\widehat{\boldsymbol{\theta}}_{\mathrm{Oracle}}$, and the third condition follows from condition (C1) and the fact that the Hessian matrix of the log-likelihood function with respect to $(\boldsymbol{\Gamma}, \boldsymbol{\Psi}, \boldsymbol{\Sigma})$ is negative definite for large enough n (as established in the proof of Lemma 1). To verify the second condition, let $u_n = n^{\tau+\eta+\kappa-1}$ for some $\eta < 1-\kappa$, and let

$$\mathcal{S}_n = \Big\{ \Big| n^{1/2} \frac{\partial}{\partial \beta_j} \mathbb{P}_n \ell(\boldsymbol{\theta}_0) \Big| \le u_n \quad \text{for } j = p_{1n}+1, \ldots, p_n \Big\}.$$

For $k \ge 1$ and $j = p_{1n}+1, \ldots, p_n$,

$$\begin{aligned} \mathrm{E}\Big| \frac{\partial}{\partial \beta_j} \ell(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \Big|^k &= \mathrm{E}\Big| \int Z(\boldsymbol{\alpha}_0^{\mathrm{T}} \boldsymbol{X} + \boldsymbol{\beta}_0^{\mathrm{T}} \boldsymbol{S}, \boldsymbol{\xi}_0) S_j f(\boldsymbol{S}^{(M)} \mid \boldsymbol{S}^{(O)}, Y, \boldsymbol{X}) \, \mathrm{d}\boldsymbol{S}^{(M)} \Big|^k \\ &\le \mathrm{E}| Z(\boldsymbol{\alpha}_0^{\mathrm{T}} \boldsymbol{X} + \boldsymbol{\beta}_0^{\mathrm{T}} \boldsymbol{S}, \boldsymbol{\xi}_0) S_j |^k, \end{aligned}$$

where $f(\boldsymbol{S}^{(M)} \mid \boldsymbol{S}^{(O)}, Y, \boldsymbol{X})$ is the true conditional density function of $\boldsymbol{S}^{(M)}$ given $(\boldsymbol{S}^{(O)}, Y, \boldsymbol{X})$, and the inequality follows from Jensen's inequality. By condition (C7) and the Bernstein inequality, there exists some fixed positive constants c_0 and c_1 such that

$$P\Big(\Big| n^{1/2} \frac{\partial}{\partial \beta_j} \mathbb{P}_n \ell(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \Big| > u_n \Big) \le c_0 e^{-c_1 u_n}$$

for $j = p_{1n}+1, \ldots, p_n$. Therefore, $P(\mathcal{S}_n) \ge 1 - c_0 p_n e^{-c_1 u_n} = 1 - O(e^{\kappa \log n - c_1 u_n})$, which tends to 1 because, by conditions (C3) and (C8), η can be chosen such that u_n dominates $\log n$. Under $\mathcal{S}_n$ and for $\widetilde{\boldsymbol{\theta}}$ such that the $\boldsymbol{\theta}_{\mathcal{S}}$-component of $\widetilde{\boldsymbol{\theta}}$ is close enough to its true value and the $\boldsymbol{\beta}_{\mathcal{N}}$-component of $\widetilde{\boldsymbol{\theta}}$ is zero,

$$\begin{aligned} \frac{\partial}{\partial \beta_j} \mathbb{P}_n \ell(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}} &= \frac{\partial}{\partial \beta_j} \mathbb{P}_n \ell(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + \frac{\partial^2}{\partial \beta_j \partial \boldsymbol{\theta}_{\mathcal{S}}^{\mathrm{T}}} \mathbb{P}_n \ell(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\check{\boldsymbol{\theta}}} (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\ &\le n^{-1/2} u_n + \sup_{\boldsymbol{\theta}^*} \Big\| \frac{\partial^2}{\partial \beta_j \partial \boldsymbol{\theta}_{\mathcal{S}}^{\mathrm{T}}} \mathbb{P}_n \ell(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \Big\| \end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

$$\leq n^{-1/2}u_n + \sup_{\boldsymbol{\theta}^*} \left\| \frac{\partial^2}{\partial \boldsymbol{\beta}_{\mathcal{N}} \partial \boldsymbol{\theta}_{\mathcal{S}}^{\mathrm{T}}} \mathbb{P}_n \ell(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \right\|_{2,\infty} \|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|$$

for $j = p_{1n}+1, \ldots, p_n$ and large enough n, where $\check{\boldsymbol{\theta}}$ takes some value between $\widetilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$, and the supremum is taken within the small neighborhood around $\boldsymbol{\theta}_0$ defined in condition (C8). Note that

$$\begin{aligned}
\sup_{\boldsymbol{\theta}^*} \left\| \frac{\partial^2}{\partial \boldsymbol{\beta}_{\mathcal{N}} \partial \boldsymbol{\theta}_{\mathcal{S}}^{\mathrm{T}}} \mathbb{P}_n \ell(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \right\|_{2,\infty} &= \sup_{\boldsymbol{\theta}^*} \left\| \frac{\partial^2}{\partial \boldsymbol{\beta}_{\mathcal{N}} \partial \boldsymbol{\theta}_{\mathcal{S}}^{\mathrm{T}}} \mathbb{P} \ell(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \right\|_{2,\infty} + O_p(n^{-1/2}p_n^2) \\
&= \sup_{\boldsymbol{\theta}^*} \|\boldsymbol{I}(\boldsymbol{\theta}^*)\|_{2,\infty} + o_p(1) \\
&= O_p(n^{\tau+\eta-1}),
\end{aligned}$$

where the third equality follows from condition (C8), and $\mathbb{P}$ is the true probability measure. Therefore,

$$\begin{aligned}
\max_{j>p_{1n}} \left| \lambda_n^{-1} w_j^{-1} \frac{\partial}{\partial \beta_j} \mathbb{P}_n \ell(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\mathrm{Oracle}}} \right| &\leq \max_{j>p_{1n}} O_p(\lambda_n^{-1} w_j^{-1} n^{-1/2} u_n + \lambda_n^{-1} w_j^{-1} n^{\tau+\eta-1} p_n n^{-1/2}) \\
&= O_p(\lambda_n^{-1} n^{-\tau} n^{-1/2} u_n + \lambda_n^{-1} n^{\eta-3/2} n^{\kappa}),
\end{aligned}$$

where the equality follows from condition (C3). The right-hand side of the equality above is bounded by $O_p(\lambda_n^{-1} n^{-3/2+\kappa+\eta})$, which is $o_p(1)$ by condition (C9). The desired result follows. □

References

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica* **71**, 135–171.

Bai, J. and Li, K. (2012). Statistical analysis of factor models of high dimension. *Ann. Statist.* **40**, 436–465.

Bai, J. and Liao, Y. (2016). Efficient estimation of approximate factor models via penalized maximum likelihood. *J. Econometrics* **191**, 1–18.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

Cai, T., Cai, T. T., and Zhang, A. (2016). Structured matrix completion with applications to genomic data integration. *J. Amer. Statist. Assoc.* **111**, 621–633.

Caner, M. and Han, X. (2014). Selecting the correct number of factors in approximate factor models: the large panel case with group bridge estimators. *J. Bus. Econom. Statist.* **32**, 359–374.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **39**, 1–22.

Fan, C., Prat, A., Parker, J. S., Liu, Y., Carey, L. A., Troester, M. A., and Perou, C. M. (2011). Building prognostic models for breast cancer patients using clinical variables and hundreds of gene expression signatures. *BMC Medical Genom.* **4**, 3.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348–1360.

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75**, 603–680.

Fan, J., Liu, H., and Wang, W. (2018). Large covariance estimation through elliptical factor models. *Ann. Statist.* **46**, 1383–1414.

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans. Inform. Theory* **57**, 5467–5484.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32**, 928–961.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.* **38**, 3567–3604.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

coordinate descent. *J. Stat. Softw.* **33**, 1–22.

Garcia, R. I., Ibrahim, J. G., and Zhu, H. (2010). Variable selection for regression models with missing data. *Statist. Sinica* **20**, 149–165.

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* **144**, 646–674.

Heiss, F. and Winschel, V. (2008). Likelihood approximation by numerical integration on sparse grids. *J. Econometrics* **144**, 62–80.

Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., Ng, S., Leiserson, M. D. M., Niu, B., McLellan, M. D., Uzunangelov, V., Zhang, J., Kandoth, C., Akbani, R., Shen, H., Omberg, L., Chu, A., Margolin, A. A., van't Veer, L. J., Lopez-Bigas, N., Laird, P. W., Raphael, B. J., Ding, L., Robertson, A. G., Byers, L. A., Mills, G. B., Weinstein, J. N., Van Waes, C., Chen, Z., Collisson, E. A., The Cancer Genome Atlas Research Network, Benz, C. C., Perou, C. M., and Stuart, J. M. (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929–944.

Ibrahim, J. G., Zhu, H., Garcia, R. I., and Guo, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics* **67**, 495–503.

Ibrahim, J. G., Zhu, H., and Tang, N. (2008). Model selection criteria for missing-data problems using the EM algorithm. *J. Amer. Statist. Assoc.* **103**, 1648–1658.

Jiang, J., Nguyen, T., and Rao, J. S. (2015). The E-MS algorithm: model selection with incomplete data. *J. Amer. Statist. Assoc.* **110**, 1136–1147.

Läubli, H., Stevenson, J. L., Varki, A., Varki, N. M., and Borsig, L. (2006). L-selectin facilitation of metastasis involves temporal induction of Fut7-dependent ligands at sites of tumor cell arrest. *Cancer Res.* **66**, 1536–1542.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K.,

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

and Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet* **11**, 733–739.

Liu, Q. and Pierce, D. A. (1994). A note on Gauss-Hermite quadrature. *Biometrika* **81**, 624–629.

Lock, E. F., Hoadley, K. A., Marron, J. S., and Nobel, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann. Appl. Stat.* **7**, 523–542.

Nicholson, R. I., Gee, J. M. W., and Harper, M. E. (2001). EGFR and cancer prognosis. *Eur. J. Cancer* **37**, S9–S15.

Normanno, N., De Luca, A., Bianco, C., Strizzi, L., Mancino, M., Maiello, M. R., Carotenuto, A., De Feo, G., Caponigro, F., and Salomon, D. S. (2006). Epidermal growth factor receptor (EGFR) signaling in cancer. *Gene* **366**, 2–16.

Ogawa, J., Inoue, H., and Koide, S. (1997). α-2,3-sialyltransferase type 3N and α-1,3-fucosyltransferase type VII are related to sialyl Lewis^x synthesis and patient survival from lung carcinoma. *Cancer* **79**, 1678–1685.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.

Shen, R., Olshen, A. B., and Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25**, 2906–2912.

Shen, R., Wang, S., and Mo, Q. (2013). Sparse integrative clustering of multiple omics data sets. *Ann. Appl. Stat.* **7**, 269–294.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* **39**, 1–13.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **58**, 267–288.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520–525.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42**, 1166–1202.

Wang, W., Baladandayuthapani, V., Morris, J. S., Broom, B. M., Manyam, G., and Do, K.-A. (2012). iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics* **29**, 149–159.

Wang, Y., Han, E., Xing, Q., Yan, J., Arrington, A., Wang, C., Tully, D., Kowolik, C. M., Lu, D. M., Frankel, P. H., Zhai, J., Wen, W., Horne, D., Yip, M. L. R., and Yim, J. H. (2015). Baicalein upregulates DDIT4 expression which mediates mTOR inhibition and growth inhibition in cancer cells. *Cancer Lett.* **358**, 170–179.

Wong, K. Y., Fan, C., Tanioka, M., Parker, J. S., Nobel, A. B., Zeng, D., Lin, D. Y., and Perou, C. M. (2019). I-Boost: an integrative boosting approach for predicting survival time with multiple genomics platforms. *Genome Biol.* **20**, 52.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76**, 217–242.

Zhang, D., Shen, D., and Alzheimer's Disease Neuroimaging Initiative (2012). Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *NeuroImage* **59**, 895–907.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7**, 2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418–1429.

Kin Yau Wong

Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0401

E-mail: kin-yau.wong@polyu.edu.hk

Donglin Zeng

Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA

E-mail: dzeng@email.unc.edu

D. Y. Lin

Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA

E-mail: lin@bios.unc.edu