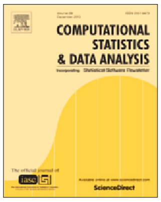
Computational Statistics and Data Analysis 157 (2021) 107127

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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



Generalized co-sparse factor regression[☆]

Aditya Mishra [a,*], Dipak K. Dey [b], Yong Chen [c], Kun Chen [b]

[a] Center for Computational Mathematics, Flatiron Institute, New York, NY 10010, USA
[b] Department of Statistics, University of Connecticut, Storrs, CT 06269, USA
[c] Division of Biostatistics, Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA

ARTICLE INFO

Article history:
Received 19 November 2019
Received in revised form 21 October 2020
Accepted 25 October 2020
Available online 3 November 2020

Keywords:
Divide-and-conquer
Integrative analysis
Multivariate learning
Singular value decomposition

ABSTRACT

Multivariate regression techniques are commonly applied to explore the associations between large numbers of outcomes and predictors. In real-world applications, the outcomes are often of mixed types, including continuous measurements, binary indicators, and counts, and the observations may also be incomplete. Building upon the recent advances in mixed-outcome modeling and sparse matrix factorization, *generalized co-sparse factor regression* (GOFAR) is proposed, which utilizes the flexible vector generalized linear model framework and encodes the outcome dependency through a sparse singular value decomposition (SSVD) of the integrated natural parameter matrix. To avoid the estimation of the notoriously difficult joint SSVD, GOFAR proposes both *sequential* and *parallel* unit-rank estimation procedures. By combining the ideas of alternating convex search and majorization–minimization, an efficient algorithm is developed to solve the sparse unit-rank problem and implemented in the R package gofar. Extensive simulation studies and two real-world applications demonstrate the effectiveness of the proposed approach.

1. Introduction

Advances in science and technology have led to exponential growth in the collection of different types of large data in various fields, including healthcare, biology, economics, and finance. Many problems of interest amount to exploring the association between multivariate outcomes/responses and multivariate predictors/features. For example, in an ongoing project of the Framingham Heart Study (Cupples et al., 2007), researchers are interested in understanding the effect of single nucleotide polymorphisms on multiple phenotypes related to cardiovascular disease. Some phenotypes are binary, depicting medical conditions, whereas others, such as cholesterol levels, are continuous. In the Longitudinal Study of Aging (LSOA) (Stanziano et al., 2010), it is of interest to understand the association between future health status (memory, depression, cognitive ability) and predictors such as demographics, past medical conditions, and daily activities. Here some outcome measurements, such as memory score, are continuous, while others are of the categorical/indicator type.

As exemplified by the aforementioned problems, the outcome variables collected in real-world studies are often of mixed types. Moreover, the data can be of large dimensionality and may contain a substantial number of missing values. It is apparent that classical multivariate linear regression (MLR) is no longer applicable, and the approach of separately

[☆] For this work, there exists supplementary materials providing all the proofs, reproducible simulation code, additional plots, tables showing model evaluation, and the application data to demonstrate model efficacy.

[*] Corresponding author.
E-mail address: amishra@flatironinstitute.org (A. Mishra).

https://doi.org/10.1016/j.csda.2020.107127
0167-9473/

regressing each response using the predictors via a generalized linear model may also perform poorly because it ignores the potential dependency of the mixed outcomes. Our main objective in this paper is thus to tackle the problem of modeling mixed and incomplete outcomes with large-scale data.

Many existing multivariate regression methods focus on continuous outcomes. Principal component regression (Jolliffe, 1982) and multivariate ridge regression (Hoerl and Kennard, 1970; Brown and Zidek, 1980) focus on tackling the problem of multicollinearity among predictors. Reduced-rank regression (Anderson, 1951; Velu and Reinsel, 2013; Bunea et al., 2011) achieves dimension reduction and information sharing by assuming that all the responses are related to a small set of latent factors. Sparse (Tibshirani, 1996) multivariate regression models (Turlach et al., 2005; Peng et al., 2010; Obozinski et al., 2011) take advantage of certain shared sparsity patterns in the association structure. Regularized multivariate models often boil down to matrix approximation problems; see, e.g., singular-value penalized models (Yuan et al., 2007; Negahban and Wainwright, 2011; Koltchinskii et al., 2011; Chen et al., 2013), and sparse matrix factorization models (Chen et al., 2012; Chen and Huang, 2012; Bunea et al., 2012; Ma and Sun, 2014; Mishra et al., 2017).

Until recently, only a handful of methods have attempted to solve the modeling challenge with non-Gaussian and mixed outcomes. Cox and Wermuth (1992) and Fitzmaurice and Laird (1995) proposed a likelihood-based approach for bivariate responses in which one variable is discrete and the other is continuous. Prentice and Zhao (1991) and Zhao et al. (1992) utilized the generalized estimating equations framework to obtain mean and covariance estimates. She (2013) and Yee and Hastie (2003) studied the reduced-rank vector generalized linear model (RR-VGLM), assuming the outcomes are of the same type and are from an exponential family distribution (Jørgensen, 1987). Recently, Luo et al. (2018) proposed *mixed-outcome reduced-rank regression* (mRRR), extending the RR-VGLM to the more realistic scenario of mixed and incomplete outcomes. However, the method only considered rank reduction, rendering it inapplicable when many redundant or irrelevant variables are present.

Building upon the recent advances in mixed-outcome modeling and sparse matrix factorization, we propose *generalized co-sparse factor regression*, which utilizes the flexible vector generalized linear model framework (She, 2013; Luo et al., 2018) and encodes the outcome dependency through an appealing sparse singular value decomposition (SVD) of the integrated natural parameter matrix. The co-sparse SVD structure in our model, i.e., the fact that both the left and the right singular vectors are sparse, implies a flexible dependency pattern between the outcomes and the predictors: on one hand, the model allows a few latent predictors to be constructed from possibly different subsets of the original predictors, and on the other hand, the model allows the responses to be associated with possibly different subsets of the predictors. The model also covers the generalized matrix completion problem under unsupervised learning. Motivated by Chen et al. (2012) and Mishra et al. (2017), we propose computationally efficient divide-and-conquer procedures to conduct model estimation. The main idea is to extract unit-rank components of the natural parameter matrix in either a *sequential* or a *parallel* way, thus avoiding the difficult joint estimation alternative. Each step solves a generalized co-sparse unit-rank estimation problem, and these problems differ only in their offset terms, which are designed to account for the effects of other non-targeted unit-rank components. Our model also allows us to deal with the missing values in the same way as in the celebrated matrix completion. To the best of our knowledge, our approach is among the first to enable both variable selection and latent factor modeling in analyzing incomplete and mixed outcomes.

The rest of the paper is organized as follows. We propose a generalized co-sparse factor regression model in Section 2. Section 3 proposes divide-and-conquer estimation procedures, which reduce the problem to a set of generalized unit-rank estimation problems; these are then studied in detail in Section 3.3. We study the large sample property of the estimator in a unit step in Section 4. Section 5 shows the effectiveness of the proposed procedures via extensive simulation studies. Two applications, one on the longitudinal study of aging and the other on sound annotation, are presented in Section 6. We provide some concluding remarks in Section 7. All the proofs are provided in Supplementary Materials.

2. Generalized co-sparse factor regression

Consider the multivariate regression setup with n instances of independent observations, forming a response/outcome matrix $\mathbf{Y} = [y_{ik}]_{n\times q} = [\mathbf{y}_1, \dots, \mathbf{y}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times q}$, a predictor/feature matrix $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times p}$, and a control variable matrix $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_n]^{\mathrm{T}} \in \mathbb{R}^{n\times p_z}$. $\mathbf{Z}$ consists of a set of variables that should always be included in the model and are thus not regularized. Depending on the application, we consider experimental input such as age or gender (factor variable) as control variables.

We assume that each of the response variables follows a distribution in the exponential-dispersion family (Jørgensen, 1987). The probability density function of the ith entry in the kth outcome, y_{ik}, is given by

$$f(y_{ik}; \theta^*_{ik}, \phi^*_k) = \exp\left\{\frac{y_{ik}\theta^*_{ik} - b_k(\theta^*_{ik})}{a_k(\phi^*_k)} + c_k(y_{ik}; \phi^*_k)\right\}, \tag{1}$$

where θ^*_{ik} is the natural parameter, $\phi^*_k \in \mathbb{R}^+$ is the dispersion parameter, and $\{a_k(\cdot), b_k(\cdot), c_k(\cdot)\}$ are functions determined by the specific distribution; see Table 1 in Supplementary Materials for more details on some of the standard distributions in the exponential family, e.g., Gaussian, Poisson and Bernoulli. We collectively denote the natural parameters of $\mathbf{Y}$ by $\boldsymbol{\Theta}^* = [\theta^*_{ik}]_{n\times q} \in \mathbb{R}^{n\times q}$ and the dispersion parameters by $\boldsymbol{\Phi}^* = \mathrm{diag}[a_1(\phi^*_1), \dots, a_q(\phi^*_q)]$. Let $g_k = (b'_k)^{-1}$ be the canonical link function. Consequently, $\mathbb{E}(y_{ik}) = b'_k(\theta^*_{ik}) = g_k^{-1}(\theta^*_{ik})$, where $b'_k(\cdot)$ denotes the derivative function of $b_k(\cdot)$.

We model the natural parameter matrix $\boldsymbol{\Theta}^*$ as

$$\boldsymbol{\Theta}(\mathbf{C}^*, \boldsymbol{\beta}^*, \mathbf{O}) = \mathbf{O} + \mathbf{Z}\boldsymbol{\beta}^* + \mathbf{X}\mathbf{C}^*, \tag{2}$$

where $\mathbf{O} = [o_{ik}]_{n\times q} \in \mathbb{R}^{n\times q}$ is a fixed offset term, $\mathbf{C}^* = [\mathbf{c}_1^*, \dots, \mathbf{c}_q^*] \in \mathbb{R}^{p\times q}$ is the coefficient matrix corresponding to the predictors, and $\boldsymbol{\beta} = [\boldsymbol{\beta}_1^*, \dots, \boldsymbol{\beta}_q^*] \in \mathbb{R}^{p_z\times q}$ is the coefficient matrix corresponding to the control variables. The intercept is included by taking the first column of $\mathbf{Z}$ to be $\mathbf{1}_n$, the $n \times 1$ vector of ones. For simplicity, we may write $\boldsymbol{\Theta}(\mathbf{C}^*, \boldsymbol{\beta}^*, \mathbf{O})$ as $\boldsymbol{\Theta}^*$ if no confusion arises.

To proceed further, we define some notations. The kth column of $\boldsymbol{\Theta}^*$ is denoted $\boldsymbol{\Theta}^*_{.k}$, and consequently $\mathbf{b}_k(\boldsymbol{\Theta}^*_{.k}) = [b_k(\theta^*_{ik}), \dots, b_k(\theta^*_{nk})]^{\mathrm{T}}$. The element-wise derivative vector of $\mathbf{b}_k(\boldsymbol{\Theta}^*_{.k})$ is $\mathbf{b}'_k(\boldsymbol{\Theta}^*_{.k}) = [b'_k(\theta^*_{ik}), \dots, b'_k(\theta^*_{nk})]^{\mathrm{T}}$. We then define

$$\mathbf{B}(\boldsymbol{\Theta}^*) = [\mathbf{b}_1(\boldsymbol{\Theta}^*_{.1}), \dots, \mathbf{b}_q(\boldsymbol{\Theta}^*_{.q})], \quad \mathbf{B}'(\boldsymbol{\Theta}^*) = [\mathbf{b}'_1(\boldsymbol{\Theta}^*_{.1}), \dots, \mathbf{b}'_q(\boldsymbol{\Theta}^*_{.q})]. \tag{3}$$

Similarly, $\mathbf{B}''(\boldsymbol{\Theta}^*)$ denotes the second-order derivative of $\mathbf{B}(\boldsymbol{\Theta}^*)$.

We assume the outcomes are conditionally independent given $\mathbf{X}$ and $\mathbf{Z}$. Then the joint negative log-likelihood function is given by

$$\mathcal{L}(\boldsymbol{\Theta}^*, \boldsymbol{\Phi}^*) = -\sum_{i=1}^{n}\sum_{k=1}^{q} \ell_k(\theta^*_{ik}, \phi^*_k), \tag{4}$$

where $\ell_k(\theta^*_{ik}, \phi^*_k) = \log f(y_{ik}; \theta^*_{ik}, \phi^*_k)$. Using the definition from (3), a convenient representation of (4) is given by

$$\mathcal{L}(\boldsymbol{\Theta}^*, \boldsymbol{\Phi}^*) = -\operatorname{tr}(\mathbf{Y}^{\mathrm{T}}\boldsymbol{\Theta}^*\boldsymbol{\Phi}^{-1*}) + \operatorname{tr}(\mathbf{J}^{\mathrm{T}}\mathbf{B}(\boldsymbol{\Theta}^*)\boldsymbol{\Phi}^{-1*}), \tag{5}$$

where $\mathbf{J} = \mathbf{1}_{n\times q}$ and $\operatorname{tr}(\mathbf{A})$ is the *trace* of a square matrix $\mathbf{A}$. In the presence of missing entries in $\mathbf{Y}$, let us define an index set of the observed outcomes as

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

and denote the projection of $\mathbf{Y}$ onto $\boldsymbol{\Omega}$ by $\widetilde{\mathbf{Y}} = \mathcal{P}_{\boldsymbol{\Omega}}(\mathbf{Y})$, where $\tilde{y}_{ik} = y_{ik}$ for any $(i, k) \in \boldsymbol{\Omega}$ and $\tilde{y}_{ik} = 0$ otherwise. Accordingly, the negative log-likelihood function with incomplete data is given by

$$\mathcal{L}(\boldsymbol{\Theta}^*, \boldsymbol{\Phi}^*) = -\operatorname{tr}(\widetilde{\mathbf{Y}}^{\mathrm{T}}\boldsymbol{\Theta}^*\boldsymbol{\Phi}^{-1*}) + \operatorname{tr}(\widetilde{\mathbf{J}}^{\mathrm{T}}\mathbf{B}(\boldsymbol{\Theta}^*)\boldsymbol{\Phi}^{-1*}),$$

where $\widetilde{\mathbf{J}} = \mathcal{P}_{\boldsymbol{\Omega}}(\mathbf{J})$. Henceforth, we mainly focus on the complete data case (5) when presenting our proposed model, as the extension to the missing data case by and large only requires replacing $\mathbf{Y}$ by $\widetilde{\mathbf{Y}}$ and $\mathbf{J}$ by $\widetilde{\mathbf{J}}$.

Without imposing additional structural assumptions on the parameters, maximum likelihood estimation, i.e., minimizing $\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\Phi})$ with respect to $\{\mathbf{C}, \boldsymbol{\beta}, \boldsymbol{\Phi}\}$ for $\boldsymbol{\Theta} = \mathbf{O} + \mathbf{X}\mathbf{C} + \mathbf{Z}\boldsymbol{\beta}$, does not work in high-dimensional settings. The marginal modeling approach, i.e., the fitting of a univariate generalized linear model (uGLM) (or its regularized version) for each individual response, would ignore the dependency among the outcomes. The mixed reduced rank regression (mRRR) (Luo et al., 2018) imposes a rank constraint on $\mathbf{C}$, but its usage is limited as it does not explore variable selection.

We assume that the regression association is driven by a few latent factors, each of which is constructed from a possibly different subset of the predictors, and, moreover, that each response may be associated with a possibly different subset of the latent factors. To be specific, this amounts to assuming a *co-sparse* SVD of $\mathbf{C}^*$ (Mishra et al., 2017), i.e., we decompose $\mathbf{C}^*$ as

$$\mathbf{C}^* = \mathbf{U}^*\mathbf{D}^*\mathbf{V}^{*\mathrm{T}}, \qquad \text{s.t.} \qquad \mathbf{U}^{*\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{U}^*/n = \mathbf{V}^{*\mathrm{T}}\mathbf{V}^* = \mathbf{I}_{r^*}, \tag{6}$$

where both the left singular vector matrix $\mathbf{U}^* = [\mathbf{u}^*_1, \dots, \mathbf{u}^*_{r^*}] \in \mathbb{R}^{p\times r^*}$ and the right singular vector matrix $\mathbf{V}^* = [\mathbf{v}^*_1, \dots, \mathbf{v}^*_{r^*}] \in \mathbb{R}^{q\times r}$ are assumed to be *sparse*, and $\mathbf{D} = \operatorname{diag}\{d^*_1, \dots, d_{r^*}\} \in \mathbb{R}^{r^*\times r^*}$ is the diagonal matrix with the nonzero singular values on its diagonal. The orthogonality constraints ensuring identifiability suggest that the sample latent factors, i.e., $(1/\sqrt{n})\mathbf{X}\mathbf{u}^*_k$ for $k = 1, \dots, r^*$, are uncorrelated with each other, and the strength of the association between the latent factors and the multivariate response $\mathbf{Y}$ is denoted by the singular values $\{d^*_1, \dots, d_{r^*}\}$. Fig. 1 shows a diagram of the proposed model structure. We thus term the proposed model **G**eneralized **co**-sparse **fa**ctor **r**egression (GOFAR).

3. Divide-and-conquer estimation procedures

Rather than jointly estimating all the sparse singular vectors simultaneously, which may necessarily involve identifiability constraints such as orthogonality in optimization (Uematsu et al., 2019), we take a divide-and-conquer approach. The main idea is to extract the unit-rank components of $\mathbf{XC}$ one by one, in either a sequential or a parallel way. In this way, we are able to divide the main task into a set of simpler unit-rank problems, which we then conquer in Section 3.3.

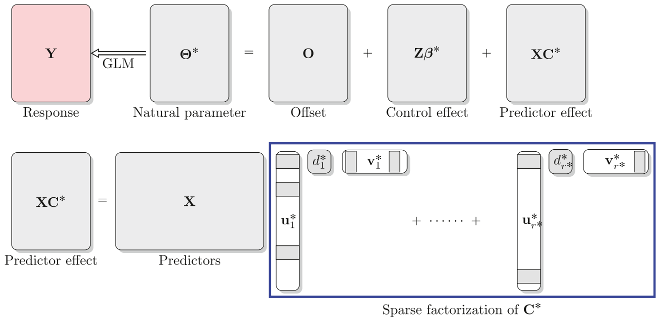


Fig. 1. GOFAR: Generalized co-sparse factor regression, modeling a multivariate mixed response matrix **Y** using a predictor matrix **X** with sparse singular vector components of the low-rank coefficient matrix $\mathbf{C}^*$.

3.1. *Sequential approach*

Motivated by Mishra et al. (2017), we propose to sequentially extract the unit-rank components of **C**, i.e., $(d_k, \mathbf{u}_k, \mathbf{v}_k)$, for $k = 1, \ldots, r$. The resulting method is termed generalized co-sparse factor regression via sequential extraction (GOFAR(S)).

Algorithm 1 and Fig. 2 summarize the computation procedure. In step $k = 1$, we conduct the following generalized co-sparse unit-rank estimation (G-CURE),

$$(\hat{d}_1, \widehat{\mathbf{u}}_1, \widehat{\mathbf{v}}_1, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Phi}}) \equiv \underset{\mathbf{u},\mathbf{d},\mathbf{v},\boldsymbol{\beta},\boldsymbol{\Phi}}{\arg\min} \ \mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\Phi}) + \rho(\mathbf{C}; \lambda), \qquad (7)$$
$$\text{s.t.} \quad \mathbf{C} = d\mathbf{u}\mathbf{v}^{\mathrm{T}},\ \mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{u}/n = \mathbf{v}^{\mathrm{T}}\mathbf{v} = 1,\ \boldsymbol{\Theta} = \boldsymbol{\Theta}(\mathbf{C}, \boldsymbol{\beta}, \mathbf{O}^{(1)}),$$

where $\mathbf{O}^{(1)} = \mathbf{O}$ (the original offset matrix), and $\rho(\mathbf{C}; \lambda)$ is a sparsity-inducing penalty function with tuning parameter λ. We discuss the formulation of $\rho(\mathbf{C}; \lambda)$ in Section 3.3.1 and the selection of tuning parameter λ in Section 3.3.4. To streamline the presentation, for now let us assume that we are able to solve G-CURE and select the tuning parameter λ suitably. Denote the produced unit-rank solution of **C** as $\widehat{\mathbf{C}}_1 = \widehat{d}_1\widehat{\mathbf{u}}_1\widehat{\mathbf{v}}_1^{\mathrm{T}}$.

In the subsequent steps, i.e., for $k = 2, \ldots, r$, we repeat G-CURE each time with an updated offset term,

$$\mathbf{O}^{(k)} = \mathbf{O} + \mathbf{X}\sum_{i=2}^{k}\widehat{\mathbf{C}}_{i-1}. \qquad (8)$$

In general, the G-CURE problem in the kth step, for $k = 1, \ldots, r$, can be expressed as

$$(\hat{d}_k, \widehat{\mathbf{u}}_k, \widehat{\mathbf{v}}_k, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Phi}}) \equiv \underset{\mathbf{u},\mathbf{d},\mathbf{v},\boldsymbol{\beta},\boldsymbol{\Phi}}{\arg\min} \ \mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\Phi}) + \rho(\mathbf{C}; \lambda), \qquad (9)$$
$$\text{s.t.} \quad \mathbf{C} = d\mathbf{u}\mathbf{v}^{\mathrm{T}},\ \mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{u}/n = \mathbf{v}^{\mathrm{T}}\mathbf{v} = 1,\ \boldsymbol{\Theta} = \boldsymbol{\Theta}(\mathbf{C}, \boldsymbol{\beta}, \mathbf{O}^{(k)}).$$

We remark that the low-dimensional parameters $\boldsymbol{\beta}$ and $\boldsymbol{\Phi}$ are re-estimated at the intermediate steps, and their final estimates are obtained from the last step.

The rationale of the proposed procedure can be traced back to the power method for computing SVD. In each step, through the construction of the offset term, the regression effects from the previous steps are adjusted or "deflated" in order to enable G-CURE to target a new unit-rank component. The procedure terminates after a pre-specified number of steps or when $\widehat{d}_k$ is estimated to be zero.

Algorithm 1 Generalized Co-Sparse Factor Regression via Sequential Extraction

```
Initialize: $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\Phi}^{(0)}$, and set the maximum number of steps $r \geq 1$, e.g., an upper bound of rank($\mathbf{C}$).
for $k \leftarrow 1$ to $r$ do
    (1) Update offset: $\mathbf{O}^{(k)} = \mathbf{O} + \mathbf{X}\sum_{i=2}^{k}\widehat{\mathbf{C}}_{i-1}$
    (2) G-CURE with tuning (see Section 3.3):
    $(\widehat{d}_k, \widehat{\mathbf{u}}_k, \widehat{\mathbf{v}}_k, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Phi}})$ = G-CURE($\mathbf{C}, \boldsymbol{\beta}, \boldsymbol{\Phi}; \mathbf{Y}, \mathbf{X}, \mathbf{O}^{(k)}, \rho$), and $\widehat{\mathbf{C}}_k = \widehat{d}_k\widehat{\mathbf{u}}_k\widehat{\mathbf{v}}_k{}^{\mathrm{T}}$.
    if $\widehat{d}_k = 0$ then
        Set $\widehat{r} = k$; $k \leftarrow r$;
    end if
end for
return $\widehat{\mathbf{C}} = \sum_{k=1}^{\widehat{r}}\widehat{\mathbf{C}}_k$, $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\Phi}}$.
```

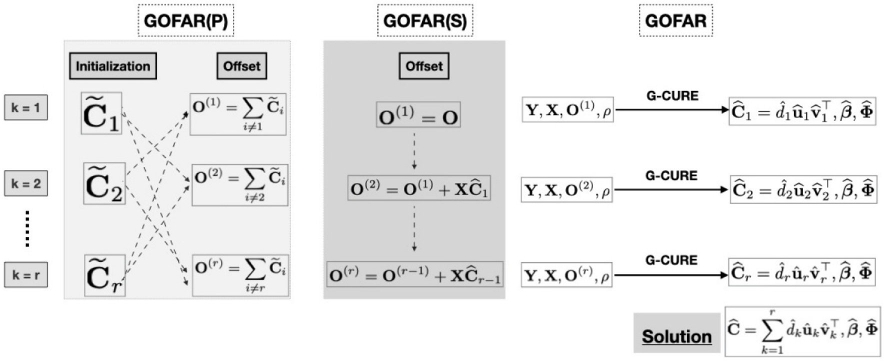

GOFAR(P)
GOFAR(S)
GOFAR
Initialization
Offset
Offset
k = 1
k = 2
k = r
$\widetilde{\mathbf{C}}_1$
$\widetilde{\mathbf{C}}_2$
$\widetilde{\mathbf{C}}_r$
$\mathbf{O}^{(1)} = \sum_{i \neq 1} \widetilde{\mathbf{C}}_i$
$\mathbf{O}^{(2)} = \sum_{i \neq 2} \widetilde{\mathbf{C}}_i$
$\mathbf{O}^{(r)} = \sum_{i \neq r} \widetilde{\mathbf{C}}_i$
$\mathbf{O}^{(1)} = \mathbf{O}$
$\mathbf{O}^{(2)} = \mathbf{O}^{(1)} + \mathbf{X}\widehat{\mathbf{C}}_1$
$\mathbf{O}^{(r)} = \mathbf{O}^{(r-1)} + \mathbf{X}\widehat{\mathbf{C}}_{r-1}$
$\mathbf{Y}, \mathbf{X}, \mathbf{O}^{(1)}, \rho$
$\mathbf{Y}, \mathbf{X}, \mathbf{O}^{(2)}, \rho$
$\mathbf{Y}, \mathbf{X}, \mathbf{O}^{(r)}, \rho$
G-CURE
G-CURE
G-CURE
$\widehat{\mathbf{C}}_1 = \hat{d}_1\hat{\mathbf{u}}_1\hat{\mathbf{v}}_1^{\mathrm{T}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Phi}}$
$\widehat{\mathbf{C}}_2 = \hat{d}_2\hat{\mathbf{u}}_2\hat{\mathbf{v}}_2^{\mathrm{T}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Phi}}$
$\widehat{\mathbf{C}}_r = \hat{d}_r\hat{\mathbf{u}}_r\hat{\mathbf{v}}_r^{\mathrm{T}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Phi}}$
Solution
$\widehat{\mathbf{C}} = \sum_{k=1}^{r} \hat{d}_k\hat{\mathbf{u}}_k\hat{\mathbf{v}}_k^{\mathrm{T}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Phi}}$

Fig. 2. Estimation procedure for the generalized co-sparse factor regression (GOFAR) via sequential (GOFAR(S)) and parallel (GOFAR(P)) extraction.

3.2. Parallel extraction

When the true rank of $\mathbf{C}$ is moderate or high, the above sequential extraction procedure may be time consuming. This motivates us to also consider generalized co-sparse factor regression via parallel extraction (GOFAR(P)), in which the construction of the offset terms for targeting different unit-rank components is based on some computationally efficient initial estimator of $\mathbf{C}$.

We summarize the GOFAR(P) procedure in Algorithm 2 and Fig. 2. Given a desired rank r, we first obtain an initial estimate of $\mathbf{C}$, denoted $\widetilde{\mathbf{C}}$, by solving an initialization problem denoted G-INIT($\mathbf{C}$, $\boldsymbol{\beta}$, $\boldsymbol{\Phi}$; $\mathbf{Y}$, $\mathbf{X}$, $\mathbf{O}$, r); see Section 1.3 of Supplementary Materials for more details. The initial estimates of the unit-rank components are then computed from the SVD of the regression components $\mathbf{X}\widetilde{\mathbf{C}}$,

$$\widetilde{\mathbf{C}} = \sum_{k=1}^{r} \widetilde{\mathbf{C}}_k = \widetilde{\mathbf{U}}\widetilde{\mathbf{D}}\widetilde{\mathbf{V}}^{\mathrm{T}}, \qquad \text{s.t. } \widetilde{\mathbf{U}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\widetilde{\mathbf{U}}/n = \widetilde{\mathbf{V}}^{\mathrm{T}}\widetilde{\mathbf{V}} = \mathbf{I}_r,$$

where $\widetilde{\mathbf{U}} = [\widetilde{\mathbf{u}}_1, \dots, \widetilde{\mathbf{u}}_r] \in \mathbb{R}^{p \times r}$, $\widetilde{\mathbf{V}} = [\widetilde{\mathbf{v}}_1, \dots, \widetilde{\mathbf{v}}_r] \in \mathbb{R}^{q \times r}$, $\widetilde{\mathbf{D}} = \mathrm{diag}[\widetilde{d}_1, \dots, \widetilde{d}_r] \in \mathbb{R}^{r \times r}$, and $\widetilde{\mathbf{C}}_k = \widetilde{d}_k\widetilde{\mathbf{u}}_k\widetilde{\mathbf{v}}_k{}^{\mathrm{T}}$.

The required offset terms for targeting different components are computed based on $\widetilde{\mathbf{C}}$ as

$$\widetilde{\mathbf{O}}^{(k)} = \mathbf{O} + \mathbf{X}\sum_{i \neq k} \widetilde{\mathbf{C}}_i, \qquad k = 1, \dots, r. \tag{10}$$

Then, the problems G-CURE($\mathbf{C}$, $\boldsymbol{\beta}$, $\boldsymbol{\Phi}$; $\mathbf{Y}$, $\mathbf{X}$, $\widetilde{\mathbf{O}}^{(k)}$), $k = 1, \dots, r$, can be solved in parallel. GOFAR(P) obtains the final estimate of $\boldsymbol{\beta}$ and $\boldsymbol{\Phi}$ from the output of the rth (last) parallel procedure.

It is clear that the quality of the initial estimator directly affects both the computational efficiency and the model accuracy of GOFAR(P). In practice, we recommend using either the mixed-outcome reduced-rank estimator proposed by Luo et al. (2018) when the model dimension is moderate or the lasso estimator when the model dimension is very high.

Algorithm 2 Generalized Co-sparse Factor Regression via Parallel Extraction

```
Initialization:
  (1) Solve $\{\widetilde{\mathbf{D}}, \widetilde{\mathbf{U}}, \widetilde{\mathbf{V}}, \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\Phi}}\}$ = G-INIT($\mathbf{C}$, $\boldsymbol{\beta}$, $\boldsymbol{\Phi}$; $\mathbf{Y}$, $\mathbf{X}$, $\mathbf{O}$, $r$) and obtain $\widetilde{\mathbf{C}}_k$ (Section 1.3 of Supplementary Materials).
  (2) Compute offsets: $\widetilde{\mathbf{O}}^{(k)} = \mathbf{O} + \mathbf{X}\sum_{i \neq k} \widetilde{\mathbf{C}}_i$, for $k = 1, \dots, r$.

for $k \leftarrow 1$ to $r$ do
    G-CURE with tuning (see Section 3.3):                                                  ⎫
    $(\widehat{d}_k, \widehat{\mathbf{u}}_k, \widehat{\mathbf{v}}_k, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Phi}})$ = G-CURE($\mathbf{C}$, $\boldsymbol{\beta}$, $\boldsymbol{\Phi}$; $\mathbf{Y}$, $\mathbf{X}$, $\widetilde{\mathbf{O}}^{(k)}$, $\rho$);   ⎬ in parallel
    $\widehat{\mathbf{C}}_k = \widehat{d}_k\widehat{\mathbf{u}}_k\widehat{\mathbf{v}}_k{}^{\mathrm{T}}$.                                          ⎭
end for
return $\widehat{\mathbf{C}} = \sum_{k=1}^{r} \widehat{\mathbf{C}}_k$, $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\Phi}}$.
```

3.3. Generalized co-sparse unit-rank estimation

3.3.1. Choice of penalty function

We denote the generic G-CURE problem as

G-CURE($\mathbf{C}$, $\boldsymbol{\beta}$, $\boldsymbol{\Phi}$; $\mathbf{Y}$, $\mathbf{X}$, $\mathbf{O}$, ρ).

First, we discuss the choice of the penalty function. In this work we use the elastic net penalty and its adaptive version (Zou and Hastie, 2005; Zou and Zhang, 2009; Mishra et al., 2017), i.e., for the kth step,

$$\rho(\mathbf{C};\lambda)=\rho(\mathbf{C};\mathbf{W},\lambda,\alpha)=\alpha\lambda\|\mathbf{W}\circ\mathbf{C}\|_1+(1-\alpha)\lambda\|\mathbf{C}\|_F^2. \tag{11}$$

Here $\|\cdot\|_1$ denotes the ℓ_1 norm, the operator "$\circ$" stands for the Hadamard product, $\mathbf{W}=[w_{ij}]_{p\times q}$ is a pre-specified weighting matrix, λ is a tuning parameter controlling the overall amount of regularization, and $\alpha\in(0,1)$ controls the relative weights between the two penalty terms. Several other penalties, such as the lasso ($\alpha=1,\gamma=0$), the adaptive lasso ($\alpha=1,\gamma>0$), and the elastic net ($0<\alpha<1,\ \gamma=0$), are its special cases.

In the kth step of GOFAR(S) or GOFAR(P), we let $\mathbf{W}_k=|\widetilde{\mathbf{C}}_k|^{-\gamma}$, where $\gamma=1$ and $\widetilde{\mathbf{C}}_k=\widetilde{d}_k\widetilde{\mathbf{u}}_k\widetilde{\mathbf{v}}_k^{\mathrm{T}}$ is an initial estimate of $\mathbf{C}_k$. As such, $w_{ijk}=w_k^{(d)}w_{ik}^{(u)}w_{jk}^{(v)}$, with

$$w_k^{(d)}=|\widetilde{d}_k|^{-\gamma},\ \mathbf{w}_k^{(u)}=[w_{1k}^{(u)},\ldots,w_{pk}^{(u)}]^{\mathrm{T}}=|\widetilde{\mathbf{u}}_k|^{-\gamma},\ \mathbf{w}_k^{(v)}=[w_{1k}^{(v)},\ldots,w_{qk}^{(v)}]^{\mathrm{T}}=|\widetilde{\mathbf{v}}_k|^{-\gamma}. \tag{12}$$

Compared to lasso, a small amount of ridge penalty in the elastic net allows correlated predictors to be in or out of the model together, thereby improving the convexity of the problem and enhancing the stability of optimization (Zou and Hastie, 2005; Mishra et al., 2017); in our work, we fix $\alpha=0.95$ and write $\rho(\mathbf{C};\mathbf{W},\lambda,\alpha)=\rho(\mathbf{C};\mathbf{W},\lambda)$ for simplicity. Now we express G-CURE($\mathbf{C},\boldsymbol{\beta},\boldsymbol{\Phi};\mathbf{Y},\mathbf{X},\mathbf{O},\rho$) as

$$\begin{aligned}(\hat{d},\widehat{\mathbf{u}},\widehat{\mathbf{v}},\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Phi}})\equiv&\underset{\mathbf{u},\mathbf{d},\mathbf{v},\boldsymbol{\beta},\boldsymbol{\Phi}}{\arg\min}\ \Big\{F_\lambda(d,\mathbf{u},\mathbf{v},\boldsymbol{\beta},\boldsymbol{\Phi})=\mathcal{L}(\boldsymbol{\Theta},\boldsymbol{\Phi})+\rho(\mathbf{C};\mathbf{W},\lambda)\Big\},\\ &\text{s.t.}\quad \mathbf{C}=d\mathbf{u}\mathbf{v}^{\mathrm{T}},\ \mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{u}/n=\mathbf{v}^{\mathrm{T}}\mathbf{v}=1,\ \boldsymbol{\Theta}=\boldsymbol{\Theta}(\mathbf{C},\boldsymbol{\beta},\mathbf{O}).\end{aligned} \tag{13}$$

3.3.2. A blockwise coordinate descent algorithm

To solve the problem in (13), we propose an iterative algorithm that cycles through a $\mathbf{u}$-step, a $\mathbf{v}$-step, a $\boldsymbol{\beta}$-step and a $\boldsymbol{\Phi}$-step to update the unknown parameters in blocks of $(\mathbf{u},d)$, $(\mathbf{v},d)$, $\boldsymbol{\beta}$ and $\boldsymbol{\Phi}$, respectively, until convergence. Below we describe each of these steps in detail.

$\mathbf{u}$-*step.* For fixed $\{\mathbf{v},\boldsymbol{\beta},\boldsymbol{\Phi}\}$ with $\mathbf{v}^{\mathrm{T}}\mathbf{v}=1$, we rewrite the objective function (13) in terms of the product variable $\check{\mathbf{u}}=d\mathbf{u}$ to avoid the quadratic constraints. For simplicity, we write $\boldsymbol{\Theta}(\mathbf{C},\boldsymbol{\beta},\mathbf{O})$ as $\boldsymbol{\Theta}(\mathbf{C})$. Motivated by She (2012) and Luo et al. (2018), we construct a *convex* surrogate of the objective function (13) with respect to $\check{\mathbf{u}}$ as follows,

$$\begin{aligned}&G_\lambda(\mathbf{a};\check{\mathbf{u}})=\mathcal{L}(\boldsymbol{\Theta}(\mathbf{a}\mathbf{v}^{\mathrm{T}}),\boldsymbol{\Phi})+\mathrm{tr}(\{\mathbf{B}'(\boldsymbol{\Theta}(\check{\mathbf{u}}\mathbf{v}^{\mathrm{T}}))\}^{\mathrm{T}}\mathbf{X}(\mathbf{a}-\check{\mathbf{u}})\mathbf{v}^{\mathrm{T}}\boldsymbol{\Phi}^{-1})-\\ &\mathrm{tr}(\mathbf{J}^{\mathrm{T}}[\mathbf{B}(\boldsymbol{\Theta}(\mathbf{a}\mathbf{v}^{\mathrm{T}}))-\mathbf{B}(\boldsymbol{\Theta}(\check{\mathbf{u}}\mathbf{v}^{\mathrm{T}}))]\boldsymbol{\Phi}^{-1})+\frac{s_u}{2}\|\mathbf{a}-\check{\mathbf{u}}\|_2^2+\rho(\mathbf{a}\mathbf{v}^{\mathrm{T}};\mathbf{W},\lambda)\\ &=\frac{s_u}{2}\|\mathbf{a}-\check{\mathbf{u}}-\frac{\mathbf{X}^{\mathrm{T}}}{s_u}[\mathbf{Y}-\mathbf{B}'(\boldsymbol{\Theta}(\check{\mathbf{u}}\mathbf{v}^{\mathrm{T}}))]\boldsymbol{\Phi}^{-1}\mathbf{v}\|_2^2+\rho(\mathbf{a}\mathbf{v}^{\mathrm{T}};\mathbf{W},\lambda)+\mathrm{const},\end{aligned} \tag{14}$$

where s_u is a scaling factor for the $\mathbf{u}$-step and "const" represents any remaining term that does not depend on the optimization variables; in this case, it is $\mathbf{a}\in\mathbb{R}^p$. It is easy to verify that $G_\lambda(\check{\mathbf{u}};\check{\mathbf{u}})=F_\lambda(d,\mathbf{u},\mathbf{v},\boldsymbol{\beta},\boldsymbol{\Phi})$. We show in the convergence analysis (see Section 3.3.3) that F is majorized by $G_\lambda(\mathbf{a};\check{\mathbf{u}})$ with appropriate scaling factor s_u. The problem of minimizing $G_\lambda(\mathbf{a};\check{\mathbf{u}})$ is separable in each entry of the vector $\mathbf{a}$. Hence, following Zou and Hastie (2005), the unique optimal solution is given by

$$\begin{aligned}\widehat{\mathbf{a}}=\boldsymbol{\mathcal{S}}(&\check{\mathbf{u}}+\mathbf{X}^{\mathrm{T}}[\mathbf{Y}-\mathbf{B}'(\boldsymbol{\Theta}(\check{\mathbf{u}}\mathbf{v}^{\mathrm{T}}/s_u))]\boldsymbol{\Phi}^{-1}\mathbf{v};\\ &\alpha\lambda\mathbf{v}^{\mathrm{T}}\mathbf{w}^{(v)}w^{(d)}\mathbf{w}^{(u)}/s_u)/\{1+2\lambda(1-\alpha)\|\mathbf{v}\|_2^2/s_u\},\end{aligned} \tag{15}$$

where $\boldsymbol{\mathcal{S}}(\mathbf{t};\tilde{\lambda})=\mathrm{sign}(\mathbf{t})(|\mathbf{t}|-\tilde{\lambda})_+$ is the elementwise soft-thresholding operator on any $\mathbf{t}\in\mathbb{R}^p$. Now, using the equality constraint, i.e., $\|\mathbf{X}\mathbf{u}\|_2=\sqrt{n}$, we can retrieve the individual estimates of $(d,\mathbf{u})$ from $\widehat{\mathbf{a}}$.

$\mathbf{v}$-*step.* As in the $\mathbf{u}$-step, we rewrite the objective function (13) in terms of the product $\check{\mathbf{v}}=d\mathbf{v}$. A *convex* surrogate that majorizes the objective function (13) with respect to $\check{\mathbf{v}}$ is constructed as

$$\begin{aligned}&H_\lambda(\mathbf{b};\check{\mathbf{v}})=\mathcal{L}(\boldsymbol{\Theta}(\mathbf{u}\mathbf{b}^{\mathrm{T}}),\boldsymbol{\Phi})+\mathrm{tr}(\{\mathbf{B}'(\boldsymbol{\Theta}(\mathbf{u}\check{\mathbf{v}}^{\mathrm{T}}))\}^{\mathrm{T}}\mathbf{X}\mathbf{u}(\mathbf{b}-\check{\mathbf{v}})^{\mathrm{T}}\boldsymbol{\Phi}^{-1})-\\ &\quad\mathrm{tr}(\mathbf{J}^{\mathrm{T}}[\mathbf{B}(\boldsymbol{\Theta}(\mathbf{u}\mathbf{b}^{\mathrm{T}}))-\mathbf{B}(\boldsymbol{\Theta}(\mathbf{u}\check{\mathbf{v}}^{\mathrm{T}}))]\boldsymbol{\Phi}^{-1})+\frac{s_v}{2}\|\mathbf{b}-\check{\mathbf{v}}\|_2^2+\rho(\mathbf{u}\mathbf{b}^{\mathrm{T}};\mathbf{W},\lambda)\\ &=\frac{s_v}{2}\|\mathbf{b}-\check{\mathbf{v}}-\boldsymbol{\Phi}^{-1}[\mathbf{Y}-\mathbf{B}'(\boldsymbol{\Theta}(\mathbf{u}\check{\mathbf{v}}^{\mathrm{T}}))]^{\mathrm{T}}\frac{\mathbf{X}}{s_v}\mathbf{u}\|_2^2+\rho(\mathbf{u}\mathbf{b}^{\mathrm{T}};\mathbf{W},\lambda)+\mathrm{const},\end{aligned} \tag{16}$$

where s_v is a scaling factor for the $\mathbf{v}$-step and $\mathbf{b}\in\mathbb{R}^q$ is the optimization variable. Following the $\mathbf{u}$-step, the unique optimal solution minimizing $H_\lambda(\mathbf{b};\check{\mathbf{v}})$ is given by

$$\begin{aligned}&\widehat{\mathbf{b}}=\boldsymbol{\mathcal{S}}(\check{\mathbf{v}}+\boldsymbol{\Phi}^{-1}[\mathbf{Y}-\mathbf{B}'(\boldsymbol{\Theta}(\mathbf{u}\check{\mathbf{v}}^{\mathrm{T}}))]^{\mathrm{T}}\mathbf{X}\mathbf{u}/s_v;\\ &\ \alpha\lambda\mathbf{u}^{\mathrm{T}}\mathbf{w}^{(u)}w^{(d)}\mathbf{w}^{(v)}/s_v)/\{1+2\lambda_{(1-\alpha)}\|\mathbf{u}\|_2^2/s_v\}.\end{aligned} \tag{17}$$

Again, we retrieve the estimates of $(d,\mathbf{v})$ from the equality constraint $\mathbf{v}^{\mathrm{T}}\mathbf{v}=1$.

$\boldsymbol{\beta}$-step. For fixed $\mathbf{C}$ and $\boldsymbol{\Phi}$, denote $\boldsymbol{\Theta}(\boldsymbol{\beta}) = \boldsymbol{\Theta}(\mathbf{C}, \boldsymbol{\beta}, \mathbf{O})$. We construct a *convex* surrogate that majorizes the objective function (13) with respect to $\boldsymbol{\beta}$ as

$$\begin{aligned} K(\boldsymbol{\alpha}; \boldsymbol{\beta}) =& \mathcal{L}(\boldsymbol{\Theta}(\boldsymbol{\alpha}), \boldsymbol{\Phi}) + \frac{s_\beta}{2}\|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_2^2 + \mathrm{tr}(\{\mathbf{B}'(\boldsymbol{\Theta}(\boldsymbol{\beta}))\}^{\mathrm{T}}\mathbf{Z}(\boldsymbol{\alpha} - \boldsymbol{\beta})\boldsymbol{\Phi}^{-1}) - \\ & \mathrm{tr}(\mathbf{J}^{\mathrm{T}}[\mathbf{B}(\boldsymbol{\Theta}(\boldsymbol{\alpha})) - \mathbf{B}(\boldsymbol{\Theta}(\boldsymbol{\beta}))]\boldsymbol{\Phi}^{-1}) \\ =& \frac{s_\beta}{2}\|\boldsymbol{\alpha} - \boldsymbol{\beta} - \frac{\mathbf{Z}^{\mathrm{T}}}{s_\beta}\{\mathbf{Y} - \mathbf{B}'(\boldsymbol{\Theta}(\boldsymbol{\beta}))\}\boldsymbol{\Phi}^{-1}\|_F^2 + \mathrm{const}, \end{aligned} \tag{18}$$

where s_β is a scaling factor for the $\boldsymbol{\beta}$-step.

A globally optimal solution minimizing $K(\boldsymbol{\alpha}; \boldsymbol{\beta})$ is given by

$$\widehat{\boldsymbol{\alpha}} = \boldsymbol{\beta} + \mathbf{Z}^{\mathrm{T}}\{\mathbf{Y} - \mathbf{B}'(\boldsymbol{\Theta}(\boldsymbol{\beta}))\}\boldsymbol{\Phi}^{-1}/s_\beta. \tag{19}$$

$\boldsymbol{\Phi}$-step. For fixed $\mathbf{C}$ and $\boldsymbol{\beta}$, we update $\boldsymbol{\Phi}$ by minimizing the negative log-likelihood function with respect to $\boldsymbol{\Phi}$, which can be obtained by a standard algorithm such as Newton–Raphson (R Core Team, 2019).

The proposed G-CURE algorithm is summarized in Algorithm 3.

Algorithm 3 Generalized Co-Sparse Unit-Rank Estimation

```
Given: $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W}, \mathbf{O}, \kappa_0, \lambda, \alpha$.
Initialize $\mathbf{u}^{(0)} = \widetilde{\mathbf{u}}, \mathbf{v}^{(0)} = \widetilde{\mathbf{v}}, d^{(0)} = \widetilde{d}, \boldsymbol{\beta}^{(0)} = \widetilde{\boldsymbol{\beta}}, \boldsymbol{\Phi}^{(0)} = \widetilde{\boldsymbol{\Phi}}$. Set $t \leftarrow 0$.
repeat
    Set $s_u = \kappa_0\|\mathbf{X}\|^2/\varphi, s_\beta = \kappa_0\|\mathbf{Z}\|^2/\varphi, s_v = n\kappa_0/\varphi$ where $\varphi = \min(\boldsymbol{\Phi}^{(t)})$.
    (1) $\mathbf{u}$-step: Set $\check{\mathbf{u}} = d^{(t)}\mathbf{u}^{(t)}$ and $\mathbf{v} = \mathbf{v}^{(t)}$. Update $\check{\mathbf{u}}^{(t+1)}$ using Eq. (15). Recover block variable $(\tilde{d}^{(t+1)}, \mathbf{u}^{(t+1)})$ using equality constraint in Eq. (13).
    (2) $\mathbf{v}$-step: Set $\check{\mathbf{v}} = \tilde{d}^{(t+1)}\mathbf{v}^{(t)}$ and $\mathbf{u} = \mathbf{u}^{(t+1)}$. Update $\check{\mathbf{v}}^{(t+1)}$ using Eq. (17). Recover block variable $(d^{(t+1)}, \mathbf{v}^{(t+1)})$ using equality constraint in Eq. (13).
    (3) $\boldsymbol{\beta}$-step: Update $\boldsymbol{\beta}^{(t+1)}$ using Eq. (19).
    (4) $\boldsymbol{\Phi}$-step: $\boldsymbol{\Phi}^{(t+1)} = \arg\max_{\boldsymbol{\Phi}} \mathcal{L}(\boldsymbol{\Theta}(\mathbf{C}^{(t+1)}, \boldsymbol{\beta}^{(t+1)}), \boldsymbol{\Phi})$.
    $t \leftarrow t + 1$.
until convergence, e.g., the relative $\ell_2$ change in parameters is less than $\epsilon = 10^{-6}$.
return $\widehat{\mathbf{u}}, \widehat{d}, \widehat{\mathbf{v}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Phi}}$.
```

3.3.3. Monotone descending property

In Algorithm 3, we use several *convex* surrogates of the objective function in order to deal with the general form of the loss function. We show that the procedure can ensure that the objective function is monotone descending with the scaling factors s_u, s_v and s_β.

We mainly consider mixed outcomes of Gaussian, Bernoulli, and Poisson distributions as examples. To conduct a formal convergence analysis, let us denote the parameter estimates in the tth step as $\{\mathbf{u}^{(t)}, d^{(t)}, \mathbf{v}^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\Phi}^{(t)}\}$. From Algorithm 3, the $\mathbf{u}$-step produces $(\tilde{d}^{(t+1)}, \mathbf{u}^{(t+1)})$, the $\mathbf{v}$-step produces $(d^{(t+1)}, \mathbf{v}^{(t+1)})$, the $\boldsymbol{\beta}$-step produces $\boldsymbol{\beta}^{(t+1)}$, and the $\boldsymbol{\Phi}$-step produces $\boldsymbol{\Phi}^{(t+1)}$.

Now, denote $\check{\mathbf{u}}^{(t+1)} = \mathbf{u}^{(t+1)}\tilde{d}^{(t+1)}$. For $\boldsymbol{\xi}_u^{(t+1)} \in \{a\check{\mathbf{u}}^{(t)}\mathbf{v}^{(t)\mathrm{T}} + (1-a)\check{\mathbf{u}}^{(t+1)}\mathbf{v}^{(t)\mathrm{T}}; 0 < a < 1\}$ and $\boldsymbol{\zeta}(\boldsymbol{\Theta}_{.k}(\boldsymbol{\xi}_u^{(t+1)}, \boldsymbol{\beta}^{(t)}), a_k(\phi_k^{(t)})) = \mathrm{diag}[\mathbf{B}''_{.k}(\boldsymbol{\Theta}_{.k}(\boldsymbol{\xi}_u^{(t+1)}, \boldsymbol{\beta}^{(t)}))]/a_k(\phi_k^{(t)})$, we define

$$\gamma_1^{(t)} = \sup_{a \in (0,1)} \|\mathbf{X}^{\mathrm{T}} \sum_{k=1}^{q} v_k^{(t)2} \boldsymbol{\zeta}(\boldsymbol{\Theta}_{.k}(\boldsymbol{\xi}_u^{(t+1)}, \boldsymbol{\beta}^{(t)}), a_k(\phi_k^{(t)}))\mathbf{X}\|.$$

Similarly, denote $\check{\mathbf{v}}^{(t+1)} = \mathbf{v}^{(t+1)}d^{(t+1)}$. Then, for $\boldsymbol{\xi}_v^{(t+1)} \in \{a\mathbf{u}^{(t)}\check{\mathbf{v}}^{(t)\mathrm{T}} + (1-a)\mathbf{u}^{(t)}\check{\mathbf{v}}^{(t+1)\mathrm{T}}; 0 < a < 1\}$ and $\boldsymbol{\zeta}(\boldsymbol{\Theta}_{.k}(\boldsymbol{\xi}_v^{(t+1)}, \boldsymbol{\beta}^{(t)}), a_k(\phi_k^{(t)})) = \mathrm{diag}[\mathbf{B}''_{.k}(\boldsymbol{\Theta}_{.k}(\boldsymbol{\xi}_v^{(t+1)}, \boldsymbol{\beta}^{(t)}))]/a_k(\phi_k^{(t)})$, we define

$$\gamma_2^{(t)} = \max_{1 \le k \le q} \sup_{a \in (0,1)} \|\mathbf{u}^{(t)\mathrm{T}}\mathbf{X}^{\mathrm{T}}\boldsymbol{\zeta}(\boldsymbol{\Theta}_{.k}(\boldsymbol{\xi}_v^{(t+1)}, \boldsymbol{\beta}^{(t)}), a_k(\phi_k^{(t)}))\mathbf{X}\mathbf{u}^{(t)}\|$$

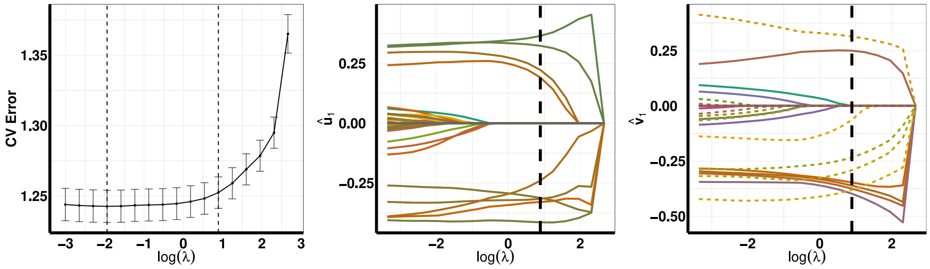
Finally, for $\boldsymbol{\xi}_\beta^{(t+1)} \in \{a\boldsymbol{\beta}^{(t)} + (1-a)\boldsymbol{\beta}^{(t+1)}; 0 < a < 1\}$ and $\mathbf{C}^{(t+1)} = d^{(t+1)}\mathbf{u}^{(t+1)}\mathbf{v}^{(t+1)\mathrm{T}}$, we define

$$\gamma_3^{(t)} = \max_{1 \le k \le q} \sup_{a \in (0,1)} \|\mathbf{Z}^{\mathrm{T}}\boldsymbol{\zeta}(\boldsymbol{\Theta}_{.k}(\mathbf{C}^{(t+1)}, \boldsymbol{\xi}_\beta^{(t+1)}), a_k(\phi_k^{(t)}))\mathbf{Z}\|.$$

Theorem 3.1. *The sequence $\{d^{(t)}, \mathbf{u}^{(t)}, \mathbf{v}^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\Phi}^{(t)}\}_{t \in \mathbb{N}}$ produced by Algorithm 3 satisfies*

$$F_\lambda(d^{(t)}, \mathbf{u}^{(t)}, \mathbf{v}^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\Phi}^{(t)}) \ge F_\lambda(d^{(t+1)}, \mathbf{u}^{(t+1)}, \mathbf{v}^{(t+1)}, \boldsymbol{\beta}^{(t+1)}, \boldsymbol{\Phi}^{(t+1)}),$$

for the scaling factors $s_u \ge \gamma_1^{(t)}$, $s_v \ge \gamma_2^{(t)}$ and $s_\beta \ge \gamma_3^{(t)}$.



(a) $\mathbf{C}$ (b) $\mathbf{C}_1$ (c) $\mathbf{C}_2$

Fig. 3. G-CURE: (a) cross-validation plot for selecting the tuning parameter λ; (b)–(c) solution paths of $d\mathbf{u}$ and $d\mathbf{v}$, respectively, in case of simulation setup I with Gaussian–Binary responses; see Table 1 for details. The dashed and continuous lines in (c) differentiate between the two types of responses.

The proof of Theorem 3.1 is relegated to Section 1.5 of Supplementary Materials. Further, we follow She (2012) to obtain the scaling factors s_u, s_v and s_β that ensure that the objective function will be monotone decreasing along the iterations. The key is to find a good upper bound of $b_k''(x)$. It is known that for Gaussian responses, $b_k''(x) = 1$ and $a_k(\phi_k) = \sigma_k^2$, and for Bernoulli responses, $b_k''(x) = e^x/(1+e^x)^2 \le 1/4$ and $a_k(\phi_k) = 1$. But for Poisson responses, $b_k''(x) = e^x$ is unbounded and $a_k(\phi_k) = 1$. Hence, in practice we choose a large enough upper bound α_p of $b_k''(x)$ empirically (default $\alpha_p = 10$).

Now, based on the above discussion, define the upper bound κ_0 for q outcomes such that $b_k''(x) \le \kappa_0$ for all $k = 1, \dots, q$. Then, at the tth step, we have $\gamma_1^{(t)} \le \kappa_0\|\mathbf{X}\|^2/\min(a_k(\phi_k^{(t)}))$; $\gamma_2^{(t)} \le \kappa_0\mathbf{u}^{(t)\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{u}^{(t)\mathrm{T}}/\min(a_k(\phi_k^{(t)})) = n\kappa_0/\min(a_k(\phi_k^{(t)}))$; and $\gamma_3^{(t)} \le \kappa_0\|\mathbf{Z}\|^2/\min(a_k(\phi_k^{(t)}))$. Hence, we set the scaling factors $s_u = \kappa_0\|\mathbf{X}\|^2/\varphi$ for the $\mathbf{u}$-step, $s_v = n\kappa_0/\varphi$ for the $\mathbf{v}$-step and $s_\beta = \kappa_0\|\mathbf{Z}\|^2/\varphi$ for the $\boldsymbol{\beta}$-step where $\varphi = \min(a_k(\phi_k^{(t)}))$.

Algorithm 3 always converges in our extensive numerical studies. Both of the estimation procedures for GOFAR are implemented, tested, validated, and made publicly available in a user-friendly R package, gofar.

3.3.4. Tuning and a toy example

Using Algorithm 3, we minimize $F_\lambda(d, \mathbf{u}, \mathbf{v}, \boldsymbol{\beta}, \boldsymbol{\Phi})$ over a range of λ values while fixing $\alpha = 0.95$ and $\gamma = 1$. The range of λ (equispaced on the log-scale), i.e., λ_{max} to λ_{min}, is chosen in order to produce a spectrum of possible sparsity patterns in $\mathbf{u}$ and $\mathbf{v}$. Specifically, $\lambda_{\max}$ is the smallest λ at which the singular value estimate is zero. In practice, we choose $\lambda_{\max} = \|\mathbf{X}^{\mathrm{T}}(\mathbf{Y} - \boldsymbol{\mu}(\mathbf{0}))\|_\infty$, and set $\lambda_{\min}$ as the fraction of $\lambda_{\max}$, i.e., $\lambda_{\min} = \lambda_{\max} \times 10^{-6}$, at which the estimated singular vectors have larger support, i.e., nonzero entries, than expected. The optimal λ can then be selected by K-fold cross-validation (Stone, 1974).

Fig. 3 shows the solution paths in simulation setup I with Gaussian–Binary responses; see Table 1 for details. The models on the solution paths are compared by the cross-validated negative log-likelihood. As with the implementation of glmnet, we suggest using the one-standard-deviation rule to select the final solution.

4. Theoretical properties

In order to focus on the large sample properties of the estimate of the unit-rank components of $\mathbf{C}$, we assume that the dispersion parameters $\boldsymbol{\Phi}$ are known. Now, without loss of generality, we set $\boldsymbol{\Phi} = \mathbf{I}$ and $\mathbf{O} = \mathbf{0}$. Using the natural parameter $\boldsymbol{\Theta}^*$ formulated in Eq. (2) and the notations defined in Eq. (3), we represent the multivariate model (1) for mixed outcomes as

$$\mathbf{Y} = \mathbf{B}'(\boldsymbol{\Theta}^*) + \mathbf{E}, \tag{20}$$

where

A1. the entries of the error $\mathbf{E} = [e_{ik}]$ are independent (σ^2, b)-sub-exponential random variables with expectation $\mathbb{E}(e_{ij}) = 0$.

In large sample theory, we let n tend to infinity with (p, q) fixed. To ensure identifiability of the parameters, we make the following assumptions on the covariates $(\mathbf{X}, \mathbf{Z})$ and the true coefficient matrix $\mathbf{C}^*$.

A2. $(1/n)\mathbf{X}^{\mathrm{T}}\mathbf{X} \overset{a.s}{\longrightarrow} \boldsymbol{\Gamma}_1$, $(1/n)\mathbf{Z}^{\mathrm{T}}\mathbf{Z} \overset{a.s}{\longrightarrow} \boldsymbol{\Gamma}_2$ and $(1/n)\mathbf{X}^{\mathrm{T}}\mathbf{Z} \overset{a.s}{\longrightarrow} \mathbf{0}$ as $n \to \infty$, where $\boldsymbol{\Gamma}_1$ and $\boldsymbol{\Gamma}_2$ are fixed, positive definite matrices.

A3. $d_1^* > \cdots > d_{r^*}^* > 0$.

To conveniently present our analysis, we allow each of the singular values d_k^* to be absorbed into the pair $(\mathbf{u}_k^*, \mathbf{v}_k^*)$ of the decomposition (6) (Chen et al., 2012). Specifically, let ℓ_k denote the index of any nonzero entry $\mathbf{v}_k^*$. Then, a uniquely identifiable reparameterization $\mathbf{C}_k^*$ is given by

$$\mathbf{C}_k^* = \mathbf{u}_k^*\mathbf{v}_k^{*\mathrm{T}}, \qquad \text{s.t.} \qquad v_{\ell_k k}^* = 1.$$

This results in $(\mathbf{u}_k^{*\mathrm{T}}\boldsymbol{\Gamma}\mathbf{u}_k^*)(\mathbf{v}_k^{*\mathrm{T}}\mathbf{v}_k^*) = d_k^*$. Consequently,

$$\begin{aligned}\mathbf{C}^* = \mathbf{U}^*\mathbf{V}^{*\mathrm{T}}, \quad & \text{s.t. } \mathbf{U}^{*\mathrm{T}}\boldsymbol{\Gamma}\mathbf{U}^* \text{ and } \mathbf{V}^{*\mathrm{T}}\mathbf{V}^* \text{ are both diagonal matrices,}\\ & \qquad\qquad v_{\ell_k k}^* = 1, k = 1, \ldots, r^*. \end{aligned} \tag{21}$$

In terms of the new parameterization, the objective function of the G-CURE optimization problem (13) is given by

$$F_k^{(n)}(\mathbf{u}, \mathbf{v}, \boldsymbol{\beta}) = \mathcal{L}(\mathbf{C}, \boldsymbol{\beta}; \mathbf{O}_k) + \rho(\mathbf{C}; \mathbf{W}_k, \lambda_k^{(n)}), \tag{22}$$

where $\mathbf{u} \in \mathbb{R}^p$, $\mathbf{v} \in \mathbb{R}^q$ with $v_{\ell_k} = 1$, $\mathbf{C} = \mathbf{u}\mathbf{v}^{\mathrm{T}}$, and the offset matrix $\mathbf{O}_k$ depends on the choice of the estimation procedure, i.e., GOFAR(S) or GOFAR(P), and mainly follows from Eqs. (8) and (10). Here $\mathbf{W}_k = [w_{ijk}]_{p\times q} = [w_{ik}w_{jk}]_{p\times q}$, where $w_{ik} = |\widetilde{u}_{ik}|^{-\gamma}$ and $w_{jk} = |\widetilde{v}_{jk}|^{-\gamma}$ for some $\gamma > 0$. The regularization parameter $\lambda_k^{(n)}$ is a function of the sample size, but $0 < \alpha \le 1$ is considered as a fixed constant. In our model formulation, $b_k''(\theta_{ik})$ corresponds to the variance of the estimate of the ikth outcome for θ_{ik}. Motivated by Luo et al. (2018), we assume that

A4. $b_k(\cdot)$ is a continuously differentiable, real-valued and strictly convex function, and the entries of the natural parameter $\boldsymbol{\Theta}$ defined in (2) satisfy

$$\min_{\substack{1\le i\le n\\ 1\le k\le q}} \inf_{\{\boldsymbol{\beta},\mathbf{C}\}} |b_k''(\theta_{ik})| \ge \gamma^l,$$

for some constant $\gamma^l > 0$.

Moreover, GOFAR(P) requires an initial estimate of the unit-rank components of the rank-r coefficient matrix $\mathbf{C}$, given by $\widetilde{\mathbf{C}}_i$ for $i = 1, \ldots, r$. We require the initial estimators to be $\sqrt{n}$-consistent, i.e.,

A5. $\|\widetilde{\mathbf{C}}_i - \mathbf{C}_i^*\| = O_p(n^{-1/2})$ for $i = 1, \ldots, r$.

This can be achieved by the unpenalized GLM estimators or the reduced-rank estimator (Velu and Reinsel, 2013; Luo et al., 2018), although these estimators do not have the desired sparse SVD structure.

Theorem 4.1. *Assume **A1–A5** hold and $\lambda_k^{(n)}/\sqrt{n} \to \lambda_k \ge 0$ as $n \to \infty$. Then the estimator $(\widehat{\mathbf{u}}_k, \widehat{\mathbf{v}}_k, \widehat{\boldsymbol{\beta}})$, from either the sequential or the parallel estimation, is $\sqrt{n}$-consistent, i.e.,*

i. $\|\widehat{\mathbf{u}}_k - \mathbf{u}_k^*\| = O_p(n^{-1/2})$, $\|\widehat{\mathbf{v}}_k - \mathbf{v}_k^*\| = O_p(n^{-1/2})$, and $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| = O_p(n^{-1/2})$ for $k = 1, \ldots, r^*$.
ii. $|\widehat{d}_k| = O_p(n^{-1/2})$ where $\widehat{d}_k = (1/n)(\widehat{\mathbf{u}}_k{}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\widehat{\mathbf{u}}_k)(\widehat{\mathbf{v}}_k{}^{\mathrm{T}}\widehat{\mathbf{v}}_k)$, for $k = r^* + 1, \ldots, r$.

Here, we have mainly followed the setup of Mishra et al. (2017) to prove the required results, the details of which are relegated to Supplementary Materials, Section 1.6. Similarly, by following Mishra et al. (2017), we can establish the selection consistency of GOFAR(S) and GOFAR(P) under assumptions **A1–A5**.

5. Simulation

5.1. Setup

We compare the estimation performance, prediction accuracy and sparsity recovery of GOFAR(S) and GOFAR(P) to those of the following modeling strategies: (a) uGLM: fit each response by the univariate sparse GLM implemented in the R package glmnet (Friedman et al., 2010); and (b) mRRR: fit by mixed-outcome reduced-rank regression (Luo et al., 2018). In addition, to show the merit of jointly learning from mixed outcomes, we also use GOFAR(S) to fit each type of responses separately; the resulting method is labeled GOFAR(S,S).

We have summarized all the simulation settings in Table 1. The setup covers scenarios with the same type of outcomes and with mixed types of outcomes. In the first scenario, the outcomes are either Gaussian (G), Bernoulli (B) or Poisson (P), whereas in the second scenario, the outcomes consist of an equal number of (a) Gaussian and Bernoulli (G–B) or (b) Gaussian and Poisson (G–P) outcomes. Moreover, setup I and setup II refer to the low-dimensional and high-dimensional simulation examples, respectively.

We set the true rank as $r^* = 3$. Denote the true coefficient matrix as $\mathbf{C}^* = \mathbf{U}^*\mathbf{D}^*\mathbf{V}^{*\mathrm{T}}$, with $\mathbf{U}^* = [\mathbf{u}_1^*, \mathbf{u}_2^*, \mathbf{u}_3^*]$, $\mathbf{V}^* = [\mathbf{v}_1^*, \mathbf{v}_2^*, \mathbf{v}_3^*]$ and $\mathbf{D}^* = s \times \mathrm{diag}[d_1^*, d_2^*, d_3^*]$. We set $d_1^* = 6$, $d_2^* = 5$, $d_3^* = 4$ and $s = 1$, except that when Poisson outcomes are present we set $s = 0.4$. The particular choice of the default value of $\alpha_p = 10$ for Poisson outcomes ensures a monotone descending objective function for the G-CURE optimization problem (13). Let $\mathrm{unif}(\mathcal{A}, b)$ denote a vector of length b whose entries are uniformly distributed on the set $\mathcal{A}$, and $\mathrm{rep}(a, b)$ denote the vector of length

Table 1
Simulation: model dimensions of all the simulation settings, including the sample size n, the number of predictors p, and the numbers $\{q_1, q_2, q_3\}$ of Gaussian (G), Bernoulli (B) and Poisson (P) outcomes, respectively.

Setup	n	p	Single-type scenario			Mixed-type scenario	
			G	B	P	G–B	G–P
I	200	100	(30,0,0)	(0,30,0)	(0,0,30)	(15,15,0)	(15,0,15)
II	200	300	(30,0,0)	(0,30,0)	(0,0,30)	(15,15,0)	(15,0,15)

Table 2
Simulation: model evaluation based on 100 replications using various performance measures (standard deviations are shown in parentheses) in Setup II with Bernoulli responses.

	Er($\mathbf{C}$)	Er($\boldsymbol{\Theta}$)	FPR	FNR	R%	r	Time (s)
	M% = 0						
GOFAR(S)	22.59 (5.15)	41.29 (6.59)	2.40 (0.98)	4.00 (2.95)	0.00 (0.00)	3.00 (0.00)	247.21 (13.82)
GOFAR(P)	26.08 (6.94)	55.10 (14.65)	6.26 (2.70)	3.30 (2.98)	0.00 (0.00)	3.00 (0.00)	49.34 (6.96)
mRRR	149.30 (12.27)	272.49 (27.92)	100.00 (0.00)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)	51.17 (0.80)
uGLM	58.11 (2.92)	120.95 (6.69)	72.56 (5.68)	1.40 (1.47)	18.17 (2.74)	25.14 (1.51)	8.05 (0.19)
	M% = 20						
GOFAR(S)	28.69 (4.96)	54.95 (7.36)	2.74 (0.95)	7.66 (4.19)	0.00 (0.00)	3.00 (0.00)	274.09 (15.10)
GOFAR(P)	38.03 (9.65)	88.77 (25.48)	8.81 (3.37)	5.63 (4.15)	0.00 (0.00)	3.00 (0.00)	54.54 (7.14)
mRRR	150.53 (24.42)	307.31 (48.92)	81.65 (20.33)	17.96 (19.89)	0.00 (0.00)	2.45 (0.61)	50.96 (0.93)
uGLM	63.93 (2.90)	140.12 (8.66)	67.57 (6.81)	2.94 (2.71)	28.70 (12.66)	24.72 (1.62)	5.90 (0.19)

b with all entries equal to a. For the single-type response scenario, we generate $\mathbf{u}_k^*$ as $\mathbf{u}_k^* = \check{\mathbf{u}}_k/\|\check{\mathbf{u}}_k\|$, where $\check{\mathbf{u}}_1 = [\text{unif}(\mathcal{A}_u, 8), \text{rep}(0, p-8)]^{\mathrm{T}}$, $\check{\mathbf{u}}_2 = [\text{rep}(0, 5), \text{unif}(\mathcal{A}_u, 9), \text{rep}(0, p-14)]^{\mathrm{T}}$, and $\check{\mathbf{u}}_3 = [\text{rep}(0, 11), \text{unif}(\mathcal{A}_u, 9), \text{rep}(0, p-20)]^{\mathrm{T}}$; and we generate $\mathbf{v}_k^*$ as $\mathbf{v}_k^* = \check{\mathbf{v}}_k/\|\check{\mathbf{v}}_k\|$, where $\check{\mathbf{v}}_1 = [\text{unif}(\mathcal{A}_v, 5), \text{rep}(0, q-5)]^{\mathrm{T}}$, $\check{\mathbf{v}}_2 = [\text{rep}(0, 5), \text{unif}(\mathcal{A}_v, 5), \text{rep}(0, q-10)]^{\mathrm{T}}$, and $\check{\mathbf{v}}_3 = [\text{rep}(0, 10), \text{unif}(\mathcal{A}_v, 5), \text{rep}(0, q-15)]^{\mathrm{T}}$. Here we set $\mathcal{A}_u = \pm 1$ and $\mathcal{A}_v = [-1, -0.3] \cup [0.3, 1]$. For the mixed-type scenario, while the $\mathbf{u}_k^*$s are generated in the same way, we set the $\mathbf{v}_k^*$s to make sure there is sufficient sharing of information among the different types of responses. Specifically, we generate $\mathbf{v}_k^*$ as $\check{\mathbf{v}}_k = [\bar{\mathbf{v}}_k, \bar{\mathbf{v}}_k]^{\mathrm{T}}$ for $k = 1, 2, 3$, where $\bar{\mathbf{v}}_1 = [\text{unif}(\mathcal{A}_u, 5), \text{rep}(0, q/2 - 5)]$, $\bar{\mathbf{v}}_2 = [\text{rep}(0, 3), \bar{v}_{14}, -\bar{v}_{15}, \text{unif}(\mathcal{A}_u, 3), \text{rep}(0, q/2 - 8)]$, and $\bar{\mathbf{v}}_3 = [\bar{v}_{11}, -\bar{v}_{12}, \text{rep}(0, 4), \bar{v}_{27}, -\bar{v}_{28}, \text{unif}(\mathcal{A}_u, 2), \text{rep}(0, q-10)]$. In all the settings, we set $\mathbf{Z} = \mathbf{1}_n$ with $\boldsymbol{\beta}^* = [\text{rep}(0.5, q)]^{\mathrm{T}}$, to include an intercept term.

The predictor matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ is generated from a multivariate normal distribution with some rotations to make sure that the latent factors $\mathbf{XU}^*/\sqrt{n}$ are orthogonal according to the proposed GOFAR model; the details can be found in Mishra et al. (2017). The dispersion parameter $a_k(\phi_k^*) = \sigma^2$ for the Gaussian outcomes is set to make the signal-to-noise ratio (SNR) equal to 0.5. (For the Binary and Poisson outcomes, $a_k(\phi_k^*) = 1$). Finally, $\mathbf{Y}$ is generated according to model (1) with $\boldsymbol{\Theta}^* = \mathbf{Z}\boldsymbol{\beta}^* + \mathbf{XC}^*$. We also consider the incomplete data setup by randomly deleting 20% of the entries in $\mathbf{Y}$ (M% = 20). The experiment under each setting is replicated 100 times.

The model estimation performance is measured by $\text{Er}(\mathbf{C}) = \|\widehat{\mathbf{C}} - \mathbf{C}^*\|_F/(pq)$ and $\text{Er}(\boldsymbol{\Theta}) = \|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F/(nq)$. The sparsity recovery is evaluated by the false positive rate (FPR) and the false negative rate (FNR), calculated by comparing the support of $(\hat{\mathbf{u}}_k, \hat{\mathbf{v}}_k)$ to that of $(\mathbf{u}_k^*, \mathbf{v}_k^*)$ for $k = 1, \ldots, r^*$. For rank recovery, we report the mean of the rank estimates and the relative percentage of signal in the $(r^* + 1)$th component and beyond, i.e., $\text{R\%} = 100(\sum_{i=r^*+1}^{\widehat{r}} \hat{d}_i^2)/(\sum_{i=1}^{\widehat{r}} \hat{d}_i^2)$; as such, R% = 0 if the rank is not over-estimated. Finally, we depict the computational complexity in terms of mean execution time.

5.2. Simulation results

Tables 2–5 report the results for the high-dimensional models in Setup II (Table 1). Figs. 4–5 show the boxplots of the estimation errors for Setups I and II. The detailed results under Setup I are relegated to Supplementary Materials, as the results under the two setups convey similar messages.

Both GOFAR(S) and GOFAR(P) consistently outperform the other methods in terms of estimation accuracy, sparsity recovery, and rank identification at the expense of reasonably manageable execution time. In particular, we observe that GOFAR methods maintain their superiority over the other competing methods for handling incomplete data; compared to the complete data counterpart, there is only a mild deterioration in the model estimator evaluation statistics. GOFAR(P) tends to have slightly better performance, which may be owing to the use of an offset that accounts for all the information of the non-targeted unit-rank components. So depending on the computational resources, one can use either of the approaches.

The superior performance of GOFAR is due to its ability to model the underlying association between multivariate responses and high-dimensional predictors through the low-rank and sparse coefficient matrix. On the other hand, the mRRR is only equipped to handle dependency through the low-rank structure. Because of this, the noise variables are

Table 3
Simulation: model evaluation based on 100 replications using various performance measures (standard deviations are shown in parentheses) in Setup II with Poisson responses.

	Er($\mathbf{C}$)	Er(Θ)	FPR	FNR	R%	r	Time (s)
	M% = 0						
GOFAR(S)	2.22 (0.60)	3.86 (0.73)	0.53 (0.50)	1.85 (2.39)	0.00 (0.00)	3.00 (0.00)	815.40 (37.25)
GOFAR(P)	2.22 (0.59)	3.97 (0.72)	6.80 (3.46)	0.91 (1.38)	0.07 (0.16)	3.59 (0.77)	188.18 (7.01)
mRRR	12.10 (0.39)	10.64 (0.59)	100.00 (0.00)	0.00 (0.00)	11.74 (2.36)	4.00 (0.00)	54.26 (0.98)
uGLM	5.93 (0.69)	10.28 (0.79)	84.65 (4.44)	0.00 (0.00)	10.46 (1.66)	25.57 (1.46)	17.03 (0.71)
	M% = 20						
GOFAR(S)	2.74 (0.67)	4.84 (0.96)	0.67 (0.51)	3.51 (2.93)	0.00 (0.00)	3.00 (0.00)	846.06 (47.99)
GOFAR(P)	3.00 (0.77)	5.18 (0.98)	9.10 (4.22)	1.21 (1.41)	1.37 (1.73)	3.69 (0.77)	197.56 (6.14)
mRRR	13.04 (0.53)	14.95 (2.47)	100.00 (0.00)	0.00 (0.00)	8.63 (6.13)	3.67 (0.61)	54.85 (1.25)
uGLM	7.22 (0.72)	13.04 (0.94)	81.50 (4.86)	1.18 (1.49)	13.14 (2.07)	25.34 (1.46)	12.32 (0.44)

Table 4
Simulation: model evaluation based on 100 replications using various performance measures (standard deviations are shown in parentheses) in Setup II with Gaussian–Bernoulli responses.

	Er($\mathbf{C}$)	Er(Θ)	FPR	FNR	R%	r	Time (s)
	M% = 0						
GOFAR(S)	20.49 (2.97)	43.34 (5.23)	0.31 (0.29)	0.88 (1.21)	0.00 (0.00)	3.00 (0.00)	129.46 (16.65)
GOFAR(P)	14.64 (4.59)	29.60 (7.83)	6.25 (3.06)	1.54 (1.96)	0.00 (0.00)	3.00 (0.00)	29.82 (4.13)
mRRR	76.17 (8.10)	164.50 (31.49)	33.37 (0.00)	67.24 (0.00)	0.00 (0.00)	1.00 (0.00)	53.74 (1.00)
uGLM	45.57 (2.54)	86.09 (5.65)	81.40 (4.66)	0.00 (0.00)	12.69 (1.52)	23.94 (1.50)	7.74 (0.16)
GOFAR(S,S)	34.14 (5.44)	80.96 (14.23)	24.15 (5.75)	6.33 (4.83)	7.27 (9.94)	6.66 (1.37)	211.37 (45.36)
	M% = 20						
GOFAR(S)	24.64 (4.11)	51.56 (6.75)	0.38 (0.27)	2.25 (1.97)	0.00 (0.00)	3.00 (0.00)	156.40 (19.51)
GOFAR(P)	20.50 (6.32)	42.40 (11.35)	11.67 (5.01)	3.61 (3.65)	0.67 (1.33)	3.39 (0.65)	34.80 (3.25)
mRRR	79.19 (6.96)	171.80 (37.15)	39.26 (12.75)	60.80 (13.92)	0.00 (0.00)	1.18 (0.38)	54.77 (1.01)
uGLM	51.89 (2.54)	102.61 (7.21)	79.26 (5.37)	0.00 (0.00)	15.85 (2.09)	24.06 (1.64)	5.64 (0.14)
GOFAR(S,S)	38.30 (4.61)	90.68 (14.64)	20.80 (5.80)	7.73 (5.53)	3.73 (1.83)	6.00 (1.36)	193.92 (46.18)

Table 5
Simulation: model evaluation based on 100 replications using various performance measures (standard deviations are shown in parentheses) in Setup II with Gaussian–Poisson responses.

	Er($\mathbf{C}$)	Er(Θ)	FPR	FNR	R%	r	Time (s)
	M% = 0						
GOFAR(S)	2.26 (0.50)	3.66 (0.64)	0.32 (0.28)	0.71 (1.06)	0.00 (0.00)	3.00 (0.00)	686.87 (23.35)
GOFAR(P)	2.00 (0.54)	3.08 (0.55)	7.75 (4.83)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)	148.44 (5.12)
mRRR	13.88 (0.63)	32.68 (2.52)	33.37 (0.00)	67.24 (0.00)	0.00 (0.00)	1.00 (0.00)	57.01 (1.19)
uGLM	5.93 (0.57)	9.31 (0.60)	87.09 (3.12)	0.00 (0.00)	10.65 (1.33)	24.03 (1.66)	12.38 (0.38)
GOFAR(S,S)	3.69 (1.28)	8.01 (3.60)	14.16 (4.76)	0.56 (1.05)	38.72 (38.17)	5.34 (0.89)	493.27 (50.71)
	M% = 20						
GOFAR(S)	2.77 (0.58)	4.58 (0.88)	0.55 (0.46)	1.39 (1.51)	0.00 (0.00)	3.00 (0.00)	678.94 (24.96)
GOFAR(P)	3.02 (0.72)	4.21 (0.78)	15.02 (6.13)	0.00 (0.00)	0.00 (0.00)	3.00 (0.00)	147.07 (5.04)
mRRR	14.81 (0.53)	35.89 (2.47)	33.37 (0.00)	67.24 (0.00)	0.00 (0.00)	1.00 (0.00)	57.14 (1.16)
uGLM	7.08 (0.52)	11.84 (0.77)	83.87 (3.61)	0.00 (0.00)	13.51 (1.73)	24.09 (1.68)	8.90 (0.29)
GOFAR(S,S)	3.64 (1.02)	6.29 (1.92)	16.46 (4.72)	0.59 (1.15)	37.48 (37.06)	5.98 (0.90)	533.82 (47.11)

all used in the estimated factors, thereby compromising the performance of the model; it may fail to identify important factors due to this limitation, which may cause rank underestimation, particularly in the mixed-type scenario. The uGLM does not explore the shared information among the outcomes, while GOFAR(S,S) does not explore the shared information among the different types of responses; so the superior performance of GOFAR over these two models further showcases the merit of integrative multivariate learning.

6. Application

6.1. Modeling of mixed outcomes from LSOA

The Longitudinal Study of Aging (LSOA) (Stanziano et al., 2010), a joint project of the National Center for Health Statistics and the National Institute on Aging, was designed to collect data measuring medical conditions, functional status, experiences and other socioeconomic dimensions of health in an aging population (70 years of age and over). The

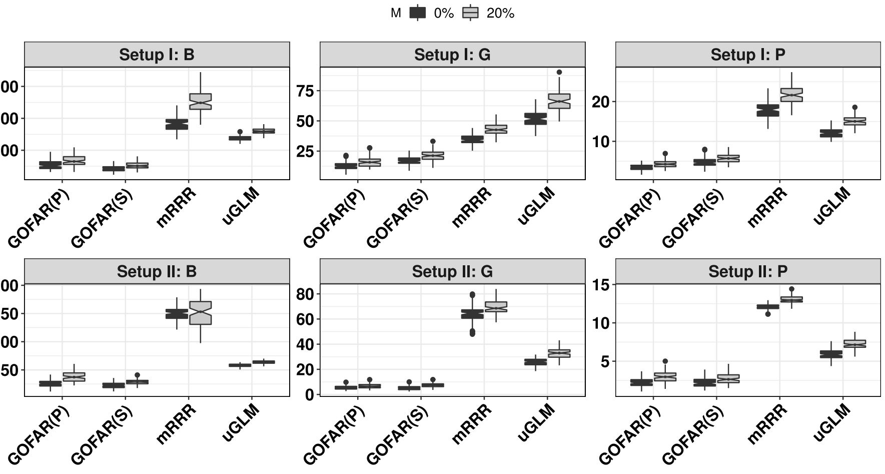


Fig. 4. Simulation: notched boxplots of the estimation error Er(**C**) for the single-type scenario under Setups I and II based on 100 replications.

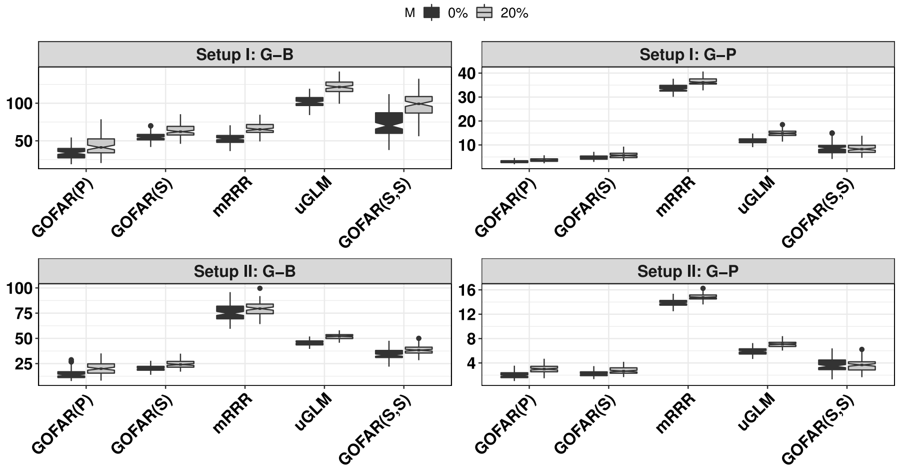


Fig. 5. Simulation: notched boxplots of the estimation error Er(**C**) for the mixed-type scenario under Setups I and II based on 100 replications.

study collected data from a large cohort of senior people in the period of 1997–1998. They were studied again between 1999–2000. Our goal is to understand the association between health-related events in the future (denoted as outcome **Y**) and health status in the past (denoted as predictor **X**) using data from $n = 3988$ subjects from this study.

The multivariate responses in **Y** include: (a) $q_1 = 3$ continuous outcomes related to overall health status, memory status and depression status; and (b) $q_2 = 41$ binary/Bernoulli outcomes related to physical conditions, medical issues, memory status, vision and hearing status, and social behavior. Our analysis considers a total of $p = 294$ predictors, constructed from the variables related to demography, family structure, daily personal care, medical history, social activity, health opinion, behavior, nutrition, health insurance, income, and assets, a majority of which are binary measurements. For simplicity, we impute missing entries in the predictors with the sample mean. GOFAR(S)/GOFAR(P) can efficiently handle missing entries in the multivariate response, so such imputations are not required for the 20.2% of entries in **Y** that are missing. Now, to determine the association between **X** and **Y**, we model the mixed outcomes jointly and apply GOFAR(S)/GOFAR(P) to obtain a low-rank and sparse estimate of the coefficient matrix. The model specifies gender and age as control variables **Z** (not penalized in the model). The parameter estimates then relate a subset of future health outcomes to a subset of past health conditions via latent factors (constructed from the subset of predictors).

On the LSOA data, GOFAR(S)/GOFAR(P) demonstrates comparable prediction performance with the advantage of producing the most parsimonious model when compared with the non-sparse method mRRR and the marginal approach uGLM. Table 6 summarizes the results from 100 replications with 75% of data selected using random sampling without replacement for training and the remaining 25% for testing. On the test data, the metric Er(G) computes the mean square error for Gaussian outcomes and the metric Er(B) computes the area under curve for binary outcomes. With the lesser number of latent factors (from r) and sufficiently sparse left and right singular vectors, GOFAR(S) produces the most parsimonious model, thus facilitating better interpretation.

Table 6
Application — LSOA: Model evaluation (standard deviations are shown in parentheses) based on prediction error of Gaussian and binary outcomes, rank estimation r and support recovery {supp(**U**) and supp(**V**)}.

Method	Er(G)	Er(B)	r	supp(U) {%}	supp(V) {%}
GOFAR(S)	0.69(0.06)	0.76(0.10)	4.45(0.65)	20(3)	43(5)
GOFAR(P)	0.72(0.06)	0.76(0.10)	4.46(0.50)	25(3)	51(5)
mRRR	0.70(0.06)	0.74(0.10)	13.15(1.75)	100(0)	100(0)
uGLM	0.68(0.06)	0.78(0.08)	41.51(0.61)	72(2)	99(0)

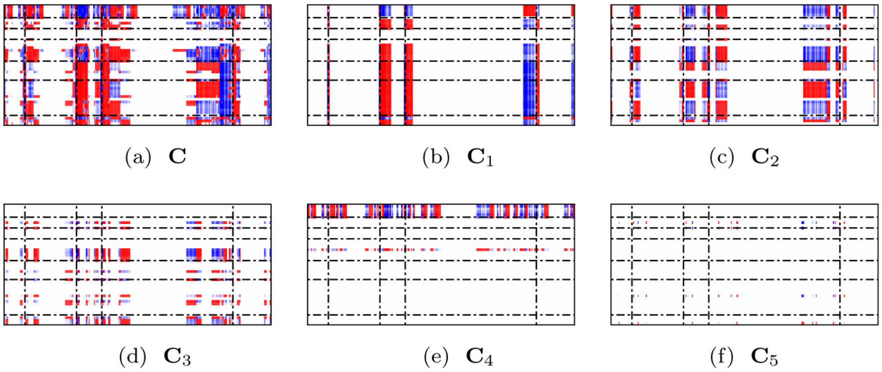

(a) $\mathbf{C}$ (b) $\mathbf{C}_1$ (c) $\mathbf{C}_2$ (d) $\mathbf{C}_3$ (e) $\mathbf{C}_4$ (f) $\mathbf{C}_5$

Fig. 6. Application — LSOA Data: The sparse estimate of the coefficient matrix $\widehat{\mathbf{C}}$ with its unit-rank components using GOFAR(S). Horizontal lines separate the response into 7 categories given by self-evaluation, fundamental daily activity, extended daily activity, medical condition, cognitive ability, sensation condition and social involvement (bottom to top). Vertical lines (left to right) separate the 294 predictors into five categories: namely, change in medical procedure since the last interview, daily activity, family status, housing condition, and prior medical condition.

Owing to the superior performance of GOFAR(S) in terms of producing the most interpretable model, we apply the procedure to the full data and obtain the parameter estimates. The GOFAR(S) approach identifies $r = 5$ subsets of outcome variables (inferred from the sparse **V**) that are associated with the predictor **X** via an equivalent number of latent factors (constructed from a subset of predictors using the sparse **U**). Fig. 6 displays the sparse estimate of the coefficient matrix **C** and its r unit-rank components. Support of the estimate of the singular vectors is given by $\text{supp}(\mathbf{U}) = \{16\%, 30\%, 34\%, 54\%, 6\%\}$ and $\text{supp}(\mathbf{V}) = \{86\%, 72\%, 34\%, 14\%, 9\%\}$. The block structure of the unit-rank components facilitates a similar interpretation, as expected from biclustering. First, latent factors constructed from a subset of predictors, mainly in the category of daily activity and prior medical conditions, determine all outcomes except cognitive ability. The latent factor clearly distinguishes social involvement outcomes from others. Apart from identifying the subset of predictors in each category, the approach finds a subgroup of the prior medical conditions affecting the outcome in the opposite way. The second latent factor helps us to identify a subgroup of the fundamental daily activity outcomes. One of the subgroups is similar to the group of outcomes related to social involvement and medical conditions. The third and fourth latent factors clearly distinguish outcomes related to social involvement from all others.

6.2. Modeling of binary outcomes from CAL500

In the second application, we consider the Computer Audition Lab 500-song (CAL500) dataset (Turnbull et al., 2007) and apply the proposed procedure to explore the underlying associations. The dataset consists of 68 audio signal characteristics from signal processing as the predictor **X**, and 174 annotations of songs by a human after listening as outcomes. The song features are mainly related to zero crossings, spectral centroid, spectral rolloff, spectral flux and Mel-Frequency Cepstral Coefficients (MFCC). On the other hand, the 174 binary outcomes from song annotations are categorized into emotions, genre, instrument, usage, vocals and song features. Some songs are annotated fewer than 20 times. We merge the disjoint sets of outcomes in a given category into one. After preprocessing, we are left with 107 binary outcomes (**Y**). Since the underlying distribution of outcomes is Bernoulli, we model the song annotations using acoustic features and apply the proposed procedure to estimate the low-rank and sparse coefficient matrix. This allows us to find subsets of song features that affect only a subset of song annotations.

As in the LSOA data analysis, we compare the parameter estimates from GOFAR(S), GOFAR(P), mRRR and uGLM, and summarize the results from 100 replicates (80% training and 20% testing) in Table 7. All the rank-constrained approaches demonstrate better prediction error performance than the marginal modeling approach (uGLM), thus proving the merit of the idea of using joint estimation to determine the underlying dependency. The prediction error performance of GOFAR(S), GOFAR(P) and mRRR are comparable, with a slight edge to mRRR. This can be attributed to the fact that the underlying system is not sufficiently sparse (see the support of **U** and **V**). We have already observed that the simulation

Table 7
Application — CAL500: Model evaluation (standard deviations are shown in parentheses) based on prediction error (PE), rank estimation r and support recovery {supp($\mathbf{U}$) and supp($\mathbf{V}$)}.

Method	PE	r	supp(U) {%}	supp(V) {%}
GOFAR(S)	0.57(0.09)	3.00(0.00)	77(4)	72(4)
GOFAR(P)	0.55(0.08)	2.71(0.65)	43(6)	46(6)
mRRR	0.58(0.10)	3.38(0.52)	100(0)	100(0)
uGLM	0.52(0.04)	20.00(0.00)	96(3)	55(5)

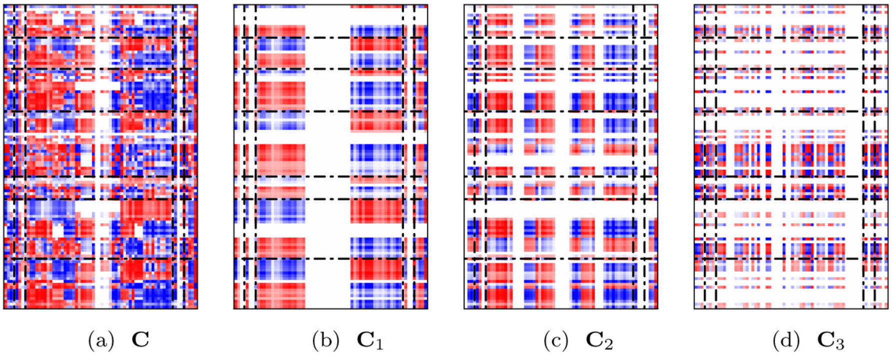

(a) $\mathbf{C}$ (b) $\mathbf{C}_1$ (c) $\mathbf{C}_2$ (d) $\mathbf{C}_3$

Fig. 7. Application — CAL500 Data: The sparse estimate of the coefficient matrix $\widehat{\mathbf{C}}$ with its unit-rank components using GOFAR(S). Vertical lines (left to right) separate 68 predictors into five categories: namely, spectral centroid, spectral flux, MFCC, spectral rolloff, and zero crossings. Horizontal lines separate the 102 response variables into seven groups: namely, emotion, genre, genre best, instrument, song, usage, and vocals (bottom to top).

results effectively demonstrate the usefulness of GOFAR(S)/GOFAR(P) in both large and high-dimensional setups where our underlying system is very sparse. Moreover, compared to the non-sparse model mRRR, GOFAR(S)/GOFAR(P) facilitates better interpretation via sparse singular vector estimates.

Again, following the LSOA data analysis, because of the better support recovery of GOFAR(S), we apply this method to analyze the full data. Fig. 7 represents the low-rank and sparse coefficient matrix $\mathbf{C}$ and its unit-rank components $\mathbf{C}_i$ for $i = 1, 2, 3$. Through the row-wise sparsity of $\mathbf{U}$, the model discards 16 predictors overall, facilitating variable selection. Support of the unit-rank components is given by supp($\mathbf{U}$) $= \{68\%, 74\%, 65\%\}$ and supp($\mathbf{V}$) $= \{81\%, 72\%, 51\%\}$. Sparsity results in block cluster representation of the unit-rank components, so we interpret it accordingly. From the first unit-rank component, we clearly identify new subgroups in the song annotation category associated with the MFCC covariates. The sign of the entries in the block matrix accordingly denotes the positive/negative associations. Among the MFCC covariates, we clearly find two separate subgroups. Blocks resulting from the second unit-rank component estimate suggest the second set of covariates (mostly disjoint from first one) that are associated with a subset of song annotations. The third unit-rank component identifies features associated with a subgroup in the outcomes related to the instrument category.

In summary, as we demonstrated in two real-world examples, the proposed GOFAR is parsimonious and effective in recovering the underlying associations through the sparse unit-rank components of the low-rank coefficient matrix.

7. Discussion

In this article, we model the mixed type of outcomes via a multivariate extension of the GLM, with each response following a distribution in the exponential dispersion family. The model encodes the response-predictor dependency through an appealing co-sparse SVD of the nature parameter matrix. We develop two estimation procedures, i.e., a sequential method, GOFAR(S) and a parallel method, GOFAR(P), to avoid the notoriously difficult joint estimation alternative.

There are many future research directions. Our model formulation (1) is restricted to outcomes from the exponential dispersion family with canonical link; it would be interesting to consider more flexible link functions and other distributional families. Theoretically, we are interested in performing non-asymptotic analysis to understand the finite sample behavior of the proposed estimators. Our approach handles missing entries in the response matrix using the same idea of matrix completion (Candès and Recht, 2009); however, it may be fruitful to further explore the role that the type of missing entry plays in parameter estimation. Moreover, it is pressing to extend our method to handle missing entries in the predictor matrix. Further, the proposed algorithms are still computationally intensive for large-scale problems; we will make the computation more scalable either by utilizing acceleration techniques in the current algorithms or by developing path-following algorithms and stagewise learning procedures (He et al., 2018; Chen et al., 2020). Finally, Algorithm 3 is a block coordinate descent optimization procedure for minimizing the nonsmooth and nonconvex objective functions $F_\lambda(d, \mathbf{u}, \mathbf{v}, \boldsymbol{\beta}, \boldsymbol{\Phi})$. This type of problem has been studied in, e.g., Gorski et al. (2007), Razaviyayn et al. (2013) and Mishra et al. (2017). In each of the sub-problems, the algorithm minimizes a *convex* surrogate that majorizes the objective function, which results in a unique and bounded solution when the elastic net penalty is used (Mishra et al., 2017). We aim to provide a detailed convergence analysis of the algorithm in our future work.

Acknowledgments

Kun Chen research is partially supported byU.S. National Science Foundation grants DMS-1613295 and IIS-1718798.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2020.107127.

References

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

Brown, P.J., Zidek, J.V., 1980. Adaptive multivariate ridge regression. Ann. Statist. 8, 64–74.

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

Bunea, F., She, Y., Wegkamp, M., 2012. Joint variable and rank selection for parsimonious estimation of high dimensional matrices. Ann. Statist. 40, 2359–2388.

Candès, E.J., Recht, B., 2009. Exact matrix completion via convex optimization. Found. Comput. Math. 9 (717).

Chen, K., Chan, K.S., Stenseth, N.C., 2012. Reduced rank stochastic regression with a sparse singular value decomposition. J. R. Stat. Soc. Ser. B Stat. Methodol. 74, 203–221.

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

Chen, K., Dong, R., Xu, W., Zheng, Z., 2020. Statistically guided divide-and-conquer for sparse factorization of large matrix. arXiv preprint arXiv:2003.07898.

Chen, L., Huang, J.Z., 2012. Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. J. Amer. Statist. Assoc. 107, 1533–1545.

Cox, D.R., Wermuth, N., 1992. Response models for mixed binary and quantitative variables. Biometrika 79, 441–461.

Cupples, L.A., Arruda, H.T., Benjamin, E.J., D'Agostino, R.B., Demissie, S., DeStefano, A.L., Dupuis, J., Falls, K.M., Fox, C.S., Gottlieb, D.J., et al., 2007. The framingham heart study 100k snp genome-wide association study resource: overview of 17 phenotype working group reports. BioMed Central Med. Genet. 8 (S1).

Fitzmaurice, G.M., Laird, N.M., 1995. Regression models for a bivariate discrete and continuous outcome with clustering. J. Amer. Statist. Assoc. 90, 845–852.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33 (1).

Gorski, J., Pfeuffer, F., Klamroth, K., 2007. Biconvex sets and optimization with biconvex functions: a survey and extensions. Math. Methods Oper. Res. 66 (3), 373–407.

He, L., Chen, K., Xu, W., Zhou, J., Wang, F., 2018. Boosted sparse and low-rank tensor regression. In: Advances in Neural Information Processing Systems. pp. 1009–1018.

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12, 55–67.

Jolliffe, I.T., 1982. A note on the use of principal components in regression. J. R. Stat. Soc. Ser. C. Appl. Stat. 31, 300–303.

Jørgensen, B., 1987. Exponential dispersion models. J. R. Stat. Soc. Ser. B Stat. Methodol. 49, 127–145.

Koltchinskii, V., Lounici, K., Tsybakov, A., 2011. Nuclear norm penalization and optimal rates for noisy low rank matrix completion. Ann. Statist. 39, 2302–2329.

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

Ma, Z., Sun, T., 2014. Adaptive sparse reduced-rank regression. arXiv preprint arXiv:1403.1922.

Mishra, A., Dey, D.K., Chen, K., 2017. Sequential co-sparse factor regression. J. Comput. Graph. Statist. 26, 814–825.

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

Obozinski, G., Wainwright, M.J., Jordan, M.I., 2011. Support union recovery in high-dimensional multivariate regression. Ann. Statist. 39, 1–47.

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D.Y., Pollack, J.R., Wang, P., 2010. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. Ann. Appl. Stat. 4 (53).

Prentice, R., Zhao, L., 1991. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. Biometrics 47 (825).

R Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/.

Razaviyayn, M., Hong, M., Luo, Z.Q., 2013. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. SIAM J. Optim. 23, 1126–1153.

She, Y., 2012. An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. Comput. Statist. Data Anal. 56, 2976–2990.

She, Y., 2013. Reduced rank multivariate generalized linear models for feature extraction. Stat. Interface 6, 197–209.

Stanziano, D.C., Whitehurst, M., Graham, P., Roos, B.A., 2010. A review of selected longitudinal studies on aging: past findings and future directions. J. Am. Geriat. Soc. 58, S292–S297.

Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. J. R. Stat. Soc. Ser. B Stat. Methodol. 36, 111–133.

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

Turlach, B.A., Venables, W.N., Wright, S.J., 2005. Simultaneous variable selection. Technometrics 47, 349–363.

Turnbull, D., Barrington, L., Torres, D., Lanckriet, G., 2007. Towards musical query-by-semantic-description using the cal500 data set. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Association for Computing Machinery, New York, NY, USA, pp. 439–446.

Uematsu, Y., Fan, Y., Chen, K., Lv, J., Lin, W., 2019. Sofar: large-scale association network learning. IEEE Trans. Inform. Theory 65, 4924–4939.

Velu, R., Reinsel, G.C., 2013. Multivariate Reduced-Rank Regression: Theory and Applications. Vol. 136. Springer Science & Business Media.

Yee, T.W., Hastie, T.J., 2003. Reduced-rank vector generalized linear models. Stat. Model. 3, 15–41.

Yuan, M., Ekici, A., Lu, Z., Monteiro, R., 2007. Dimension reduction and coefficient estimation in multivariate linear regression. J. R. Stat. Soc. Ser. B Stat. Methodol. 69, 329–346.

Zhao, L.P., Prentice, R.L., Self, S.G., 1992. Multivariate mean parameter estimation by using a partly exponential model. J. R. Stat. Soc. Ser. B Stat. Methodol. 54, 805–811.

Zou, H., Hastie, T.J., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. 67, 301–320.

Zou, H., Zhang, H.H., 2009. On the adaptive elastic-net with a diverging number of parameters. Ann. Statist. 37 (1733).