Statistics Innovation

METHOD

Open Access

https://doi.org/10.48130/stati-0026-0007

Statistics Innovation **2026**, 3: e007

Factor-augmented group effect selection with application to gene set analysis

Yihe Yang*

Department of Population and Quantitative Health Sciences, Case Western Reserve University School of Medicine, Cleveland, Ohio 44106-7078, US

* Correspondence: yxy1234@case.edu (Yang Y)

Abstract

This paper addresses the problem of group variable selection when both the number of groups and the number of variables within each group are large. We propose the factor-augmented group effect selection (FAGES) method, which simultaneously identifies important groups, provides comparable estimates of group effect sizes, and determines the directions of these effects. The key idea is to assume that a low-dimensional latent factor captures the major information within each group and to apply a variable selection penalty to these factors in order to select relevant groups and estimate their effects. We establish the consistency of both parameter estimation and model selection under moderate conditions. Simulation studies demonstrate that FAGES can reliably recover significant groups and estimate their effects, even when the working model is misspecified. In practice, FAGES can be applied to gene set analysis to identify important biological pathways and quantify their direct effects. Using head and neck squamous cell carcinoma data, we illustrate the practical utility of FAGES by detecting multiple pathways implicated in disease development.

Citation: Yang Y. 2026. Factor-augmented group effect selection with application to gene set analysis. *Statistics Innovation* 3: e007 https://doi.org/10.48130/stati-0026-0007

Introduction

Variable selection is a fundamental problem in modern statistics, particularly in high-dimensional regression, where the number of predictors can greatly exceed the sample size[1,2,3]. When covariates possess a natural grouping structure, such as sets of genetic markers belonging to the same biological pathway, the ability to select or discard groups of variables rather than individual variables becomes crucial for both interpretability and statistical efficiency. Group LASSO[4] provides an early and influential framework for grouped selection by penalizing the ℓ_2-norm of group coefficients. It has spurred a wide range of developments, including group smoothly clipped absolute deviation (group SCAD) and group minimax concave penalty (group MCP)[5], as well as bi-level selection approaches, such as the composite MCP[6], group bridge[7], and group exponential LASSO[8]. For more details on group variable selection, refer to Huang et al.[5].

Our research is motivated by the unique data structures encountered in gene set analysis (GSA). Gene set variation analysis (GSVA)[9] is a widely used unsupervised method that constructs, for each individual and each gene set, a Kolmogorov–Smirnov (KS) statistic comparing the distributions of expression levels for genes inside the set with those outside. The collection of such statistics along the ranked gene list yields an individualized KS curve that characterizes the enrichment profile of the gene set. Figure 1 shows examples of these curves for the Kyoto Encyclopedia of genes and genomes (KEGG)[10] pathway "MicroRNAs in Cancer," where each curve corresponds to one individual. In practice, GSVA summarizes each curve by a single enrichment score, which is then used as the gene-set–level feature. This representation suffers from two important limitations. First, the enrichment score is marginal in the sense that it does not adjust for dependence among gene sets, and thus, the gene sets with significant enrichment scores may rise because of correlation with truly relevant ones. Second, reducing the entire curve to a single number may fail to capture the important structural features of the enrichment pattern, potentially limiting the statistical efficiency and biological interpretability[11].

A natural remedy is to treat the collection of genes within a set, or equivalently, the curves formed by their KS statistics, as a group, and then apply group variable selection methods to automatically identify the most important gene sets. However, the existing group variable selection methods are not well suited for this setting because gene set data typically exhibit substantially more complex structures. For instance, the design matrix of a gene set may consist of individualized curves, leading to dimensions of size $n \times p$, where n is the sample size and p is the number of genes. When the group size is very large, conventional group selection methods are prone to computational difficulties such as non-convergence. Moreover, strong correlations among genes or functional-like curve structures within a set result in severe multicollinearity, violating the irrepresentable conditions required for consistent group selection[12,13]. Consequently, truly important gene sets cannot be reliably recovered under standard penalized regression. A few methods, such as the elastic net[14] and factor-adjusted regularized model selection (Farm-Select)[15], offer partial remedies by addressing multicollinearity, but they neither exploit the grouping of covariates nor accommodate complex within-group structures.

To address the challenges of high-dimensional grouped data with complex group structures, we propose factor-augmented group effect selection (FAGES), a novel group variable selection method that simultaneously identifies important groups, provides comparable estimates of group effect sizes, and determines their effect directions. The key idea is to represent each group of variables by a low-dimensional latent factor that captures its main variation, and then specify a factor-augmented regression model accounting for the additive contributions of all groups. In particular, when the grouped variables take the form of curves (e.g., individualized KS curves in GSA), the latent factor representation can be viewed as functional

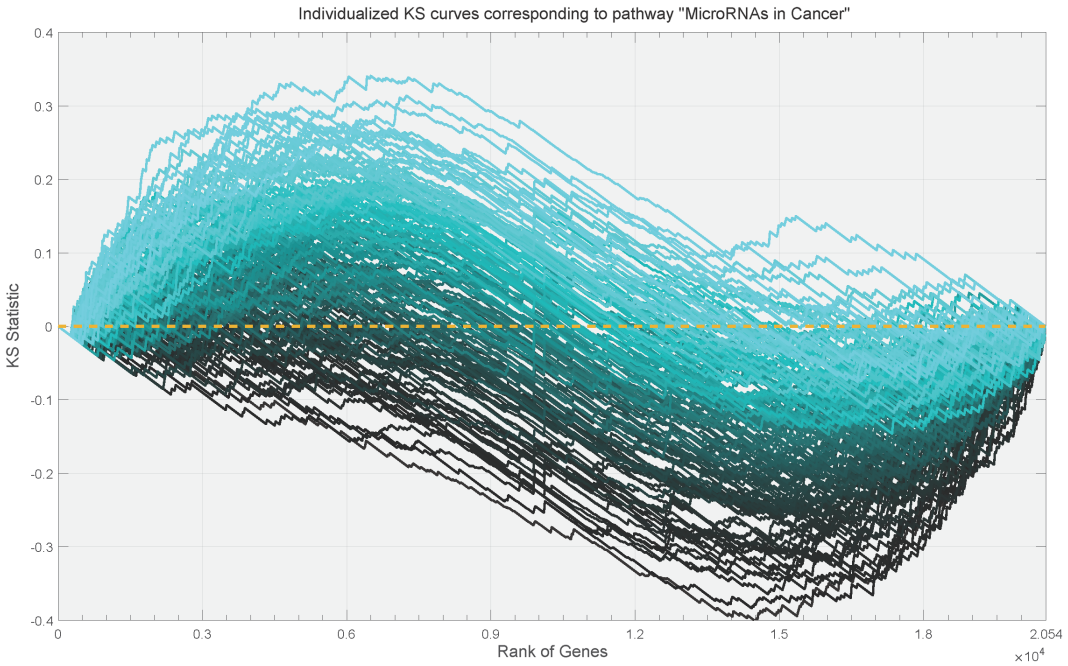

Fig. 1 Individualized KS curves corresponding to the pathway "MicroRNAs in Cancer". Each curve corresponds to an individual and is estimated from the head and neck squamous cell carcinoma data. The depth of the curve's color is determined by the value of the extreme point in the curve.

principal component analysis (FPCA)[16,17], enabling FAGES to effectively handle functional data structures. By applying a variable selection penalty, FAGES achieves both group identification and effect estimation in a unified framework. We establish the consistency of parameter estimation and group selection under mild conditions, and numerical studies demonstrate that FAGES reliably recovers relevant groups even under model misspecification. As an illustration, we apply FAGES to head and neck squamous cell carcinoma (HNSCC) transcriptomic data and detect the biological pathways that play key roles in disease development.

Materials and Methods

Notation

For a vector $\boldsymbol{a} = (a_j)_{p\times 1}$, let $||\boldsymbol{a}||_q = (\sum_{j=1}^{p}|a_j|^q)^{1/q}$ with $q \in [0, \infty]$. For a symmetric matrix $\boldsymbol{A} = (\boldsymbol{A}_{ij})_{p\times p}$, $\sigma_j(\boldsymbol{A})$ denotes the j-th largest eigenvalue of matrix $\boldsymbol{A}$, $||\mathbf{A}||_F^2 = \sum_i\sum_j A_{ij}^2$, and $||\boldsymbol{A}||_q = \max\{||\boldsymbol{A}\boldsymbol{a}||_q, ||\boldsymbol{a}||_q = 1\}$. Besides, $a_n \asymp b_n$ if there are positive constants c and C such that $c \le a_n/b_n \le C$. Notations $\{a_j\}$, and $\{\boldsymbol{A}_j\}$ represent a set of vectors $\{a_1, \ldots, a_p\}$ and a set of matrices $\{\boldsymbol{A}_1, \ldots, \boldsymbol{A}_p\}$, respectively. In addition, notation diag(a) denotes the diagonalizing operator of vector a and diag($\boldsymbol{A}_1, \ldots, \boldsymbol{A}_p$) denotes the block-diagonalizing operator of a series of matrices $\{\boldsymbol{A}_j\}$. Moreover, we define $\sigma_{\max}(\boldsymbol{A}) = \sigma_1(\boldsymbol{A})$ and $\sigma_{\min}(\boldsymbol{A}) = \sigma_p(\boldsymbol{A})$ as the largest and smallest eigenvalues of a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{p\times p}$, respectively.

Background

Approximate factor model

A multivariate variable $X_i = (X_{i1}, \ldots, X_{ip})^T$ is termed to follow a factor model if

$$\mathbf{X} = \mathbf{F}\mathbf{\Lambda}^\top + \mathbf{e}, \tag{1}$$

where, $\mathbf{X} = (X_1, \ldots, X_n)^T$ is the sample matrix of X_i, $\mathbf{F} = (F_1, \ldots, F_n)^T$ is the matrix of latent factor $F_i = (f_{i1}, \ldots, f_{iK})^T$, $\mathbf{\Lambda} = (\lambda_1, \ldots, \lambda p)^T$ is the loadings matrix with $\lambda_j = (\lambda_{j1}, \ldots, \lambda_{jK})^T$, and $\mathbf{e} = (e_1, \ldots, e_n)^T$ is the matrix of the idiosyncratic component $e_i = (e_{i1}, \ldots, e_{ip})^T$. The factor model is known as approximate factor model (AFM) if the idiosyncratic components have cross-sectional correlation, which has been verified to be effective in explaining the correlation structure of variables in econometrics[18].

Principal components analysis (PCA) and AFM are closely related[18]. Specifically, $\mathbf{F}$ and $\mathbf{\Lambda}$ can be estimated by the following restricted minimization:

$$\hat{\mathbf{F}}, \hat{\mathbf{\Lambda}} = \arg\min_{\mathbf{F},\mathbf{\Lambda}} ||\mathbf{X} - \mathbf{F}\mathbf{\Lambda}^\top||_{\mathrm{F}}^2, \quad \text{subject to } \mathbf{F}^\top\mathbf{F}/n = \mathbf{I}_K \text{ and } \mathbf{\Lambda}^\top\mathbf{\Lambda} \text{ is diagonal.} \tag{2}$$

The minimizers of the above restricted minimization are explicit and unique. Suppose the singular value decomposition $X = \sum_{s=1}^{p} d_s U_s V_s^T$, where U_s is the s-th left singular vector, V_s is the s-th right singular vector, and d_s is the s-th singular value of X. Then $\hat{\mathbf{F}} = \sqrt{n}(\boldsymbol{U}_1, \cdots, \boldsymbol{U}_K)$ and $\hat{\Lambda} = \mathbf{X}^\top\hat{\mathbf{F}}/\sqrt{n}$. That is, $\hat{\mathbf{F}}$ consists of the first K principal components (PCs) of $\mathbf{X}$.

Note that AFM is further connected to functional PCA, which is used to characterize the main pattern of the individualized trajectories around an overall mean trend function in functional data analysis[19]. In the simulation example on functional regression, we introduce how to use AFM to handle the functional data through the projection-PCA[16].

Group variable selection approaches

Multivariate variable X_i is termed to have a group structure if there is a series of sets $\{\mathcal{M}_1, \cdots, \mathcal{M}_J\}$ such that $X_i = (\boldsymbol{X}_{i,\mathcal{M}_1}^\top, \ldots, \boldsymbol{X}_{i,\mathcal{M}_J}^\top)^\top$. Multivariate GLM with grouped variables refers to

$$g(\boldsymbol{\mu}) = \mathbf{1}\beta_0 + \mathbf{X}_{\mathcal{M}_1}\boldsymbol{\beta}_{\mathcal{M}_1} + \cdots + \mathbf{X}_{\mathcal{M}_J}\boldsymbol{\beta}_{\mathcal{M}_J}, \tag{3}$$

where, $\boldsymbol{\mu} = E(y)$ and $y = (y_1, \ldots, y_n)T \in \mathbb{R}^n$ is the response that follows the exponential family distribution[20]. Here, $1 \in \mathbb{R}^n$ is the all-one vector, $\mathbf{X}_{\mathcal{M}_j} = (\boldsymbol{X}_{1,\mathcal{M}_j}, \ldots, \boldsymbol{X}_{n,\mathcal{M}_j})^\top \in \mathbb{R}^{n\times p_j}$ is the design matrix of variables in

group $\mathcal{M}_j$, and $\boldsymbol{\beta}_{\mathcal{M}_j} \in \mathbb{R}^{p_j}$ is the corresponding regression coefficient vector, where $p_j = |\mathcal{M}_j|$ and $s \in \mathcal{M}_j$. Moreover, $X = (\mathbf{1}, \mathbf{X}_{\mathcal{M}_1}, \ldots, \mathbf{X}_{\mathcal{M}_J}) \in \mathbb{R}^{n\times(1+\sum_{j=1}^{J} p_j)}$ and $\beta = (\beta_0, \boldsymbol{\beta}_{\mathcal{M}_1}^\top, \ldots, \boldsymbol{\beta}_{\mathcal{M}_J}^\top)^\top \in \mathbb{R}^{1+\sum_{j=1}^{J} p_j}$; $g(\cdot)$ is known as the canonical link function. In addition, we assume that these sets are known in advance from preliminary information. For example, the KEGG[10] database is designed according to the functional pathways of genes, while the gene ontology (GO)[21] database is made based on the ontology annotations of genes.

The regression coefficient β can be estimated through the penalized likelihood[22] shown below:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \mathcal{L}(\mathbf{X}\boldsymbol{\beta}) + n \sum_{j=1}^{J} \rho_\lambda(\|\boldsymbol{\beta}_{\mathcal{M}_j}\|) \right\}, \tag{4}$$

where, $\mathcal{L}(\boldsymbol{\eta}) = -(\mathbf{y}^\top \boldsymbol{\eta} - \mathbf{1}^\top b(\boldsymbol{\eta}))$ is the negative log-likelihood function and $b'(x)$ is the derivative of $b(x)$; η is known as the linear predictor, which is equal to $X\beta$ in this case; $b(\cdot)$ is a known function satisfying $b'(x) = g^{-1}(x)$; and $\rho_\lambda(\cdot)$ is a group variable selection penalty with a penalizing parameter λ. There are two categories of group variable selection penalties: group-level selection penalty, such as group LASSO[4], group SCAD, and group MCP[5], and bi-level selection penalty, including composite MCP (cMCP)[6], group bridge approach[7], and group exponential LASSO (GEL)[8]. For example, the group MCP is defined as:

$$\rho_\lambda^{\text{mcp}}(\|\boldsymbol{\beta}\|_2) = \lambda \int_0^{\|\boldsymbol{\beta}\|_2} \left(1 - \frac{t}{a\lambda}\right)_+ \mathrm{d}t, \tag{5}$$

where, $a > 2$ is a tuning parameter. As well, the expression of GEL is given by

$$\rho_\lambda^{\text{gel}}(\|\boldsymbol{\beta}\|_1) = \frac{\lambda^2}{a}\left\{1 - \exp\left(-\frac{a\|\boldsymbol{\beta}\|_1}{\lambda}\right)\right\}, \tag{6}$$

where, $a > 0$ is an alternative tuning parameter. In addition, a weight w_β is generally applied to adjust λ as $w_\beta\lambda$ to trade off the influence of group size. For group-level penalty, w_β is usually set as $\sqrt{\dim(\boldsymbol{\beta})}$; for bi-level penalty, w_β is often estimated as $\dim(\beta)$.

Note that whether $\rho_\lambda(\|\beta\|)$ is called group- or bi-level penalty depends on the type of norm $\|\cdot\|$ rather than its expression. The penalty is called group-level penalty if it measures the ℓ_2-norm of β, while it is termed bi-level penalty if it measures the ℓ_1-norm of β. In addition, group-level penalty can only select the important groups, whereas bi-level penalty can select important variables and groups simultaneously. For more details, refer to Huang et al.[5].

Statistical methodology of FAGES

Representation

FAGES is a novel group variable selection method that combines the AFM and group variable selection approach reviewed in the previous section. Suppose a multivariate variable with group structure, i.e., $X_i = (1, \boldsymbol{X}_{i,\mathcal{M}_1}^\top, \ldots, \boldsymbol{X}_{i,\mathcal{M}_J}^\top)^\top$; in particular, the multiple variables $\boldsymbol{X}_{i,\mathcal{M}_j}$ are usually highly correlated in the group $\mathcal{M}_j$. FAGES assumes these multiple variables to follow an AFM:

$$\mathbf{X}_{\mathcal{M}_j} = \mathbf{F}_{\mathcal{M}_j}\boldsymbol{\Lambda}_{\mathcal{M}_j}^\top + \mathbf{e}_{\mathcal{M}_j}, \tag{7}$$

where, $\mathbf{X}_{\mathcal{M}_j} = (\boldsymbol{X}_{1,\mathcal{M}_j}, \ldots, \boldsymbol{X}_{n,\mathcal{M}_j})^\top$, $\mathbf{F}_{\mathcal{M}_j} = (\boldsymbol{F}_{1,\mathcal{M}_j}, \ldots, \boldsymbol{F}_{n,\mathcal{M}_j})^\top$, $\boldsymbol{\Lambda}_{\mathcal{M}_j} = (\boldsymbol{\lambda}_{1,\mathcal{M}_j}, \ldots, \boldsymbol{\lambda}_{p_j,\mathcal{M}_j})$, and $\mathbf{e}_{\mathcal{M}_j} = (\boldsymbol{e}_{1,\mathcal{M}_j}, \ldots, \boldsymbol{e}_{n,\mathcal{M}_j})^\top$.

Since the major information of $\mathbf{X}_{\mathcal{M}_j}$ is represented by $\mathbf{F}_{\mathcal{M}_j}$, it is able to detect the important group $\mathcal{M}_j$ by looking at the significance of the corresponding latent factor $\mathbf{F}_{\mathcal{M}_j}$. Motivated by this principle, FAGES considers a new group-wise factor-augmented GLM (GF-GLM), which addresses the regression between a response y and the multiple latent factors of all groups:

$$g(\boldsymbol{\mu}) = \mathbf{1}\theta_0 + \sum_{j=1}^{J} \mathbf{F}_{\mathcal{M}_j}\boldsymbol{\theta}_{\mathcal{M}_j}. \tag{8}$$

Compared with the high-dimensional GLM (3), the grouped factor-augmented GLM discards the redundant parts $\{e_{i\mathcal{M}_j}\}$ and reduces the dimension of the model from $\sum_j^J p_j$ to $\sum_j^J K_j$. Compared with the factor-augmented regression[23], the grouped factor-augmented GLM allows the latent factors of different groups to be presented in the same regression, so that it is able to correctly prioritize the groups and make the valid inferences. In addition, if only a few latent factors of groups have significant associations with phenotype, the group LASSO and its modifications can be applied to select the non-zero $\boldsymbol{\theta}_{\mathcal{M}_j}$.

When the grouped factor-augmented GLM is misspecified, FAGES is still able to select the important groups with high precision. Specifically, consider the following model:

$$\mathrm{E}(y_i|\boldsymbol{X}_i) = g^{-1}(\eta_i^F + \eta_i^e), \tag{9}$$

Here, $\eta_i^F = \beta_0 + \sum_j \boldsymbol{F}_{i\mathcal{M}_j}^\top \boldsymbol{\theta}_j$ and $\eta_i^e = \sum_j \boldsymbol{e}_{i\mathcal{M}_j}^\top \boldsymbol{\beta}_{\mathcal{M}_j}$, which means that both the latent factors $\{\boldsymbol{F}_{i,\mathcal{M}_j}\}$ and the idiosyncratic component $\{\boldsymbol{e}_{i,\mathcal{M}_j}\}$ have effects on y_i. Using the Taylor expansion, this model reduces to

$$\mathrm{E}(y_i|\boldsymbol{X}_i) = g^{-1}(\eta_i^F) + O(|\eta_i^e|^2), \tag{10}$$

indicating that the latent factors $\{\boldsymbol{F}_{\mathcal{M}_j}\}$ can describe the main variation of y_i as long as they carry most information of $\{\boldsymbol{X}_{\mathcal{M}_j}\}$. This is why FAGES can find the important groups and accurately estimate the group effects even if the grouped factor-augmented GLM is misspecified. In the literature, such an approximation has been utilized by Hall et al.[24] to analyze the discrete functional data. They also found that a top few functional PCs (FPCs) can sufficiently describe the main variation of discrete functional data, offering the empirical confirmation of the robustness of our grouped factor-augmented generalized linear model (GLM) model.

In practice, we adopt a data-adaptive rule to select the number of factors. Specifically, for each group $\mathcal{M}_j$, let $\{\sigma_{jk}\}_{k\geq 1}$ denote the squared eigenvalues of $n^{-1}\mathbf{X}_{\mathcal{M}_j}^\top \mathbf{X}_{\mathcal{M}_j}$, truncated at $K_{\max}$ for stability. We consider three complementary criteria. The first is a gap-ratio statistic[25]:

$$z_{jk} = \frac{\sigma_{j,k-1} - \sigma_{jk}}{\sigma_{jk} - \sigma_{j,k+1}}, \qquad k = 2, \ldots, K_{\max}, \tag{11}$$

and we set $K_j^{\text{DR}} = \arg\max_{2\leq k\leq K_{\max}} z_{jk}$. The second is an eigenvalue-ratio statistic[26]:

$$r_{jk} = \frac{\sigma_{jk}}{\sigma_{j,k+1}}, \qquad k = 1, \ldots, K_{\max}, \tag{12}$$

yielding $K_j^{\text{ER}} = \arg\max_{1\leq k\leq K_{\max}} r_{jk}$. The third criterion applies a hard threshold on the spectrum of the standardized group variables[27]: letting $\tilde{\sigma}_{jk}$ be the eigenvalues of the sample correlation matrix of $\mathbf{X}_{\mathcal{M}_j}$, we define

$$K_j^{\text{ACT}} = \sum_{k=1}^{K_{\max}} \mathrm{I}\left(\tilde{\sigma}_{jk} > 1 + \sqrt{\frac{p_j}{n-1}}\right). \tag{13}$$

Finally, we take the conservative aggregation

$$K_j = \min\left\{K_j^{\text{DR}}, K_j^{\text{ER}}, K_j^{\text{ACT}}\right\}, \tag{14}$$

and estimate the group factors $\mathbf{F}_{\mathcal{M}_j}$ accordingly, which empirically stabilizes inference by avoiding overestimation of K_j across heterogeneous group sizes and within-group correlation structures. We adopt this conservative strategy because Fan et al.[27] pointed out that the primary practical risk in factor-augmented models is not underestimating the number of factors but rather including too many spurious factors, where subsequent inference may be adversely affected.

Estimation and inference

In the implementation, FAGES first estimates the latent factor of each group through the PCA and then yields the related coefficient using the penalized likelihood below:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left\{ \mathcal{L}(\hat{\mathbf{F}}\boldsymbol{\theta}) + n \sum_{j=1}^{J} \rho_{w_j\lambda}(\|\boldsymbol{\theta}_{\mathcal{M}_j}\|) \right\}, \tag{15}$$

where, $\mathcal{L}(\boldsymbol{\eta})$ is the negative log-likelihood function; $\boldsymbol{\eta} = \hat{\mathbf{F}}\boldsymbol{\theta}$ is the linear predictor; $\hat{\mathbf{F}} = (\mathbf{1}, \hat{\mathbf{F}}_{\mathcal{M}_1}, \ldots, \hat{\mathbf{F}}_{\mathcal{M}_J})$, where $\hat{\mathbf{F}}_{\mathcal{M}_j}$ is the first K_j PCs of $\mathbf{X}_{\mathcal{M}_j}$; $\|\cdot\|$ may be the ℓ_2- or the ℓ_1-norm, with respect to group- and bi-level selection approaches; and w_j is a given weight corresponding to group $\mathcal{M}_j$. This penalized likelihood reduces to the traditional penalized likelihood if the latent factors are observed or their consistent estimators are given. The block descent algorithm[28] can be used to solve (Eq. [15]) in a computationally efficient manner.

After obtaining the minimizer $\hat{\boldsymbol{\theta}}$, we propose to quantify the strength of group effect of $\mathcal{M}_j$ by using the following averaged group effect estimate:

$$\hat{z}_{\mathcal{M}_j} = \frac{1}{\sqrt{n}} \|\hat{\mathbf{F}}_{\mathcal{M}_j} \hat{\boldsymbol{\theta}}_{\mathcal{M}_j}\|_2. \tag{16}$$

This averaged group effect estimate is based on the fact that multiple latent factors may exist within a group and their directions are not identifiable under the AFM/PCA representation, rendering individual coefficient signs uninformative. By aggregating the fitted group-specific signal across samples, the averaged group effect provides a meaningful and comparable summary of the overall group contribution regardless of factor orientation.

With the same motivation, we propose to determine the related effect direction by the sign of certain well-defined correlation statistics $\widehat{\mathrm{cor}}(\hat{\mathbf{F}}_{\mathcal{M}_j}\hat{\boldsymbol{\theta}}_{\mathcal{M}_j}, \mathbf{y})$. For example, the rank correlation

$$\hat{\omega}_{\mathcal{M}_j} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{s\neq i}^{n} \mathbf{1}(\hat{\boldsymbol{F}}_{i,\mathcal{M}_j}^{\top}\hat{\boldsymbol{\theta}}_{\mathcal{M}_j} < \hat{\boldsymbol{F}}_{s,\mathcal{M}_j}^{\top}\hat{\boldsymbol{\theta}}_{\mathcal{M}_j}) \mathbf{1}(y_i < y_s) - \frac{1}{4} \tag{17}$$

has been demonstrated to be robust for modelling the dependency of two variables[29]. Other correlations like Kendall-τ correlation are also appropriate to define the sign of the cluster effect. Although the factors of a group $\hat{\mathbf{F}}_{\mathcal{M}_j}$ themselves lack a straightforward statistical interpretation, their biological significance can be elucidated by examining the signs of certain marker genes in the loading matrix $\hat{\boldsymbol{\Lambda}}_{\mathcal{M}_j}$. This approach facilitates a deeper understanding of the underlying biological implications of the factors. With this specification of the averaged group effect, FAGES is able to rank the importance of significant groups and judge the related effect directions toward phenotype.

Large sample property

Denote $\hat{\mathbf{F}}_{\mathcal{M}_j}$ as the matrix consisting of the first K_j PCs of the matrix $\mathbf{X}_{\mathcal{M}_j}$ and $\hat{\boldsymbol{\Lambda}}_{\mathcal{M}_j} = \mathbf{X}_{\mathcal{M}_j}^{\top}\hat{\mathbf{F}}_{\mathcal{M}_j}/n$, $\hat{\mathbf{V}}_{\mathcal{M}_j}$ as the $(K_j \times K_j)$ diagonal matrix consisting of the first K_j eigenvalues of matrix $\mathbf{X}_{\mathcal{M}_j}^{\top}\mathbf{X}_{\mathcal{M}_j}/(np_j)$, and $\mathbf{H}_{\mathcal{M}_j}^{\top} = \hat{\mathbf{V}}_{\mathcal{M}_j}^{-1}(\hat{\mathbf{F}}_{\mathcal{M}_j}^{\top}\mathbf{F}_{\mathcal{M}_j}/n)(\boldsymbol{\Lambda}^{\top}\boldsymbol{\Lambda}/p_j)$. Denote $\mathbf{H}_{\mathcal{M}} = \mathrm{diag}(\mathbf{H}_{\mathcal{M}_1}, \ldots, \mathbf{H}_{\mathcal{M}_{J_0}})$, $\mathbf{H}_{\mathcal{M}^c} = \mathrm{diag}(\mathbf{H}_{\mathcal{M}_{J_0+1}}, \ldots, \mathbf{H}_{\mathcal{M}_J})$, and $H = \mathrm{diag}(\mathbf{H}_{\mathcal{M}}, \mathbf{H}_{\mathcal{M}^c})$. Let $\boldsymbol{\theta}^{\star} = (\theta_0^{\star}, (\boldsymbol{\theta}_{\mathcal{M}_1}^{\star})^{\top}, \ldots, (\boldsymbol{\theta}_{\mathcal{M}_q}^{\star})^{\top})^{\top}$ be the real regression coefficient vector. Denote $\mathcal{M} = \{\mathcal{M}_j, \|\boldsymbol{\theta}_{\mathcal{M}_j}^{\star}\|_2 \neq 0\}$, $\mathcal{M}^c = \{\mathcal{M}_j, \|\boldsymbol{\theta}_{\mathcal{M}_j}^{\star}\|_2 = 0\}$, $\boldsymbol{\theta}_{\mathcal{M}}^{\star} = (\theta_0^{\star}, (\boldsymbol{\theta}_{\mathcal{M}_1}^{\star})^{\top}, \ldots, (\boldsymbol{\theta}_{\mathcal{M}_{J_0}}^{\star})^{\top})^{\top}$, and $\boldsymbol{\theta}_{\mathcal{M}^c}^{\star} = ((\boldsymbol{\theta}_{\mathcal{M}_{J_0+1}}^{\star})^{\top}, \ldots, (\boldsymbol{\theta}_{\mathcal{M}_J}^{\star})^{\top})^{\top} = 0$. Next, $\mathrm{E}(y_i) = \mu_i = b'(\boldsymbol{F}_i^{\top}\boldsymbol{\theta}^{\star})$, $\mathrm{var}(y_i) = \phi_0 b''(\mu_i) = \phi_0 b''(\boldsymbol{F}_i^{\top}\boldsymbol{\theta}^{\star})$, and $W_0 = \phi_0 \mathrm{diag}(b''(\boldsymbol{F}_1^{\top}\boldsymbol{\theta}^{\star}), \ldots, b''(\boldsymbol{F}_n^{\top}\boldsymbol{\theta}^{\star}))$. We consider the standard exponential family distribution so that the dispersion parameter $\phi_0 = 1$. Let $\boldsymbol{\epsilon} = \mathbf{y} - b'(\mathbf{F}\boldsymbol{\theta}^{\star})$ be the residual vector, $\boldsymbol{\varepsilon} = \mathbf{W}_0^{-1/2}(\mathbf{y} - b'(\mathbf{F}\boldsymbol{\theta}^{\star}))$ be the scaled residual vector, and $\boldsymbol{\delta} = (\hat{\mathbf{F}}_{\mathcal{M}} - \mathbf{F}_{\mathcal{M}}\mathbf{H}_{\mathcal{M}})\mathbf{H}_{\mathcal{M}}^{-1}\theta^{\star}$ be the estimation error of significant latent factors. In addition, denote $\mathbf{C}_{\mathcal{M}_j\mathcal{M}} = \lim_{n\to\infty} \mathbf{H}_{\mathcal{M}_j}^{\top}\mathbf{F}_{\mathcal{M}_j}^{\top}\mathbf{W}_0\mathbf{F}_{\mathcal{M}}\mathbf{H}_{\mathcal{M}}/n$ and $\mathbf{C}_{\mathcal{M}\mathcal{M}} = \lim_{n\to\infty} \mathbf{H}_{\mathcal{M}}^{\top}\hat{\mathbf{F}}_{\mathcal{M}}^{\top}\mathbf{W}_0\mathbf{F}_{\mathcal{M}}\mathbf{H}_{\mathcal{M}}/n$.

The following conditions facilitate the proofs of the theorems.

(A1) For all j, the group of variables $\boldsymbol{X}_{i,\mathcal{M}_j}$, the associated common factor $\boldsymbol{F}_{i,\mathcal{M}_j}$, the factor loading matrix $\boldsymbol{\Lambda}_{\mathcal{M}_j}$, and the idiosyncratic components $\boldsymbol{e}_{i,\mathcal{M}_j}$ satisfy the standard approximate factor condition given in the supplementary materials.

(A2) The scaled residual $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is a vector of IID variables, which satisfies that for all i, $E(\varepsilon_i) = 0$, $\mathrm{var}(\varepsilon_i) = 1$, and $\mathrm{E}(\exp(t\varepsilon_i)) \leq \exp(\tau_0 t^2/2)$ for all $t \in R$, where τ_0 is a scale parameter of the tail. Next, $\{\varepsilon_i\}$, $\{F_{i,\mathcal{M}_j}\}_{1\leq j\leq p}$, $\{e_{i,\mathcal{M}_j}\}_{1\leq j\leq p}$ are mutually independent groups. Furthermore, $\max_i E(|\varepsilon_i|^3) = O(1)$ and $n^{-\frac{3}{2}} \sum_{i=1}^{n} \|\boldsymbol{F}_{i,\mathcal{M}}^{\top}\mathbf{H}_{\mathcal{M}}\mathbf{C}_{\mathcal{M}\mathcal{M}}^{-2}\mathbf{H}_{\mathcal{M}}^{\top}\boldsymbol{F}_{i,\mathcal{M}}\|_2^3 \to 0$.

(A3) There is a positive constant c_0 such that $-c_0 < \min_i \eta_i \leq \max_i \eta_i < c_0$, where $\eta_i = \boldsymbol{F}_i^{\top}\boldsymbol{\theta}^{\star}$ for all $i \in \{1, \ldots, n\}$. There is a constant c_1 such that $|b'(\eta_i) - b'(\eta_j)| \leq c_1|\eta_i - \eta_j|$ and $|b''(\eta_i) - b''(\eta_j)| \leq c_1|\eta_i - \eta_j|$ for all $i, j \in \{1, \ldots, n\}$. Furthermore, there is a positive constant σ_0 such that $\sigma_0 < \sigma_{\min}(\mathbf{C}_{\mathcal{M}_j\mathcal{M}}^{\top}\mathbf{C}_{\mathcal{M}_j\mathcal{M}}) \leq \sigma_{\max}(\mathbf{C}_{\mathcal{M}_j\mathcal{M}}^{\top}\mathbf{C}_{\mathcal{M}_j\mathcal{M}}) < \sigma_0^{-1}$ and $\sigma_0 < \sigma_{\min}(\mathbf{C}_{\mathcal{M}\mathcal{M}}) \leq \sigma_{\max}(\mathbf{C}_{\mathcal{M}\mathcal{M}}) < \sigma_0^{-1}$ for all $j \in \{1, \ldots, J\}$.

(A4) The concave penalty $\rho_{\lambda}(\cdot)$ with concavity parameter a satisfies the condition that $\rho_{\lambda}(\|x\|)$ is increasing and concave in $\|x\| \in [0, +\infty)$ with $\rho_{\lambda}(0) = 0$, and that $\rho_{\lambda}(\|x\|)$ is differentiable in $\|x\| \in (0, +\infty)$ with $\rho'_{\lambda}(0) := \rho'_{\lambda}(0+)$. In addition, $\rho'_{\lambda}(\|x\|) \geq a_1\lambda$ for all $\|x\| \in [0, a_2\lambda]$, and $\rho'_{\lambda}(\|x\|) = o(n^{-1/2})$ for all $\|x\| \in [a\lambda, +\infty)$, for any $a > a_2$.

(A5) The dimensions of the latent factors $\{K_j\}$ are bounded, the dimensions of the variables of groups $\{p_j\}$ satisfy $p_j \asymp n$ for all j, and the weights $w_1, \ldots, w_J$ are bounded. Besides, $J_0^2 n^{-1} \to 0$ and $\lambda^{-1}\alpha_n \to 0$, where $\alpha_n = \max[(J_0/n)^{1/2}, \{\log(J)/n\}^{1/2}]$. Furthermore, for both ℓ_2- or ℓ_{∞}-norm, there is a positive constant c_0 such that $\max_{j\in\{1,\ldots,J_0\}} \|\boldsymbol{\theta}_j\| < c_0 < \infty$ and $\min_{j\in\{1,\ldots,J_0\}} \min_{s\in\mathcal{M}_j} |\boldsymbol{\theta}_{s,\mathcal{M}_j}|/\lambda \to \infty$.

Condition (A1) presents the standard conditions of factor structure given by Fan et al.[30]. Condition (A2) demonstrates that we only pay attention to exponential family distributions where the noise term ε_i is sub-Gaussian distributed[31]. In particular, the condition "$\max_i \mathrm{E}(|\varepsilon_i|^3) = O(1)$ and $n^{-\frac{3}{2}} \sum_{i=1}^{n} |\boldsymbol{F}_{i,\mathcal{M}}^{\top}\mathbf{H}_{\mathcal{M}}\mathbf{C}_{\mathcal{M}\mathcal{M}}^{-2}\mathbf{H}_{\mathcal{M}}^{\top}\boldsymbol{F}_{i,\mathcal{M}}|_2^3 \to 0$" is imposed to ensure the validity of the Lyapunov condition for asymptotic normality, which is standard in high-dimensional statistical inference; see, for example, Condition 6 in Fan & Lv[22]. Moreover, (A3) summarizes the standard conditions for the response and the Fisher information matrix of the penalized likelihood function (Eq. [15]). Condition (A4) refers to the standard conditions of group concave penalty given by Fan et al.[32]. Condition (A5) is crucial to prove the estimation consistency and selection consistency of the FAGES. Additionally, in AFM (Eq. [1]), $\mathbf{F}$ and $\boldsymbol{\Lambda}$ are not separably identifiable without the restrictions $\mathbf{F}^{\top}\mathbf{F}/n = \mathbf{I}_K$ and $\boldsymbol{\Lambda}^{\top}\boldsymbol{\Lambda}$ is a diagonal matrix, since $\mathbf{F}\boldsymbol{\Lambda}^{\top} = \mathbf{F}\mathbf{Q}^{-1}\mathbf{Q}\boldsymbol{\Lambda}$ for any invertible matrix $\mathbf{Q}$. For this problem, Bai[18] defined an identification matrix $\mathbf{H}^{\top} = (\hat{\boldsymbol{\Lambda}}^{\top}\hat{\boldsymbol{\Lambda}}/p)^{-1}(\hat{\mathbf{F}}^{\top}\mathbf{F}/n)(\boldsymbol{\Lambda}^{\top}\boldsymbol{\Lambda}/p)$, which plays the central role in the asymptotic property study of AFM. Here, we employ the group-wise identification matrices $\{\mathbf{H}_{\mathcal{M}_j}\}$ to study the asymptotic properties of FAGES.

Theorem 1 (model selection consistency) *Suppose that conditions (A1)-(A5) are satisfied. Let $\mathcal{O}_1(\hat{\boldsymbol{\theta}})$ be the event that there exists a strict local minimizer $\hat{\boldsymbol{\theta}}$ in Eq. (15) such that $\|\hat{\boldsymbol{\theta}}_{\mathcal{M}_j}\|_2 > 0$ for all $j \in \{1, \ldots, J_0\}$ and $\mathcal{O}_2(\hat{\boldsymbol{\theta}})$ be the event that $\|\hat{\boldsymbol{\theta}}_{\mathcal{M}_j}\|_2 = 0$ for all $j \in \{J_0 + 1, \ldots, J\}$. Then as $n \to \infty$, $\Pr(\mathcal{O}_1(\hat{\boldsymbol{\theta}}) \cap \mathcal{O}_2(\hat{\boldsymbol{\theta}})) \to 1$.*

Theorem 2 (parameter estimation consistency) *Suppose that conditions (A1)–(A5) are satisfied. Then, for the local minimizer* $\hat{\boldsymbol{\theta}}$ *estimated from Eq. (15),* $\|\hat{\boldsymbol{\theta}}_{\mathcal{M}} - \mathbf{H}_{\mathcal{M}}^{-1}\boldsymbol{\theta}^{\star}_{\mathcal{M}}\|_2 = O_P(\sqrt{J_0/n})$. *In addition,* $\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathcal{M}} - \mathbf{H}_{\mathcal{M}}^{-1}\boldsymbol{\theta}^{\star}_{\mathcal{M}}) \xrightarrow{D} \mathcal{N}(\mathbf{b}_{\mathcal{M}}, \mathbf{C}^{-1}_{\mathcal{M}\mathcal{M}})$, *where, the bias term* $\mathbf{b}_{\mathcal{M}} = \lim_{n\to\infty} \frac{1}{\sqrt{n}} \mathbf{C}^{-1}_{\mathcal{M}\mathcal{M}} \mathbf{H}^{\top}_{\mathcal{M}} \mathbf{F}^{\top}_{\mathcal{M}} \mathbf{W}_0 \boldsymbol{\delta}$.

Theorem 1 indicates that FAGES can achieve the model selection consistency. Theorem 2 points out the convergence rate and asymptotic normal distribution of $\hat{\boldsymbol{\theta}}_{\mathcal{M}}$, which is the same as the case where we know the sparsity pattern of $\boldsymbol{\theta}^{\star}$. In addition, we offer the asymptotic bias $\mathbf{b}_{\mathcal{M}}$ in the asymptotic normal distribution of $\hat{\boldsymbol{\theta}}_{\mathcal{M}} - \mathbf{H}_{\mathcal{M}}^{-1}\boldsymbol{\theta}^{\star}_{\mathcal{M}}$. In theory, this bias term does not vanish as $n \to \infty$ because its scale has the same order of magnitude as the variance of $\hat{\boldsymbol{\theta}}_{\mathcal{M}}$, i,e., $\|\mathbf{b}_{\mathcal{M}}\|_2^2 = O_p(J_0)$. It should be pointed out that this asymptotic bias is not estimable in theory, and our empirical analysis shows its influence is small in practice. We suggest ignoring this asymptotic bias and then make the inference using the asymptotic covariance matrix $\mathbf{C}^{-1}_{\mathcal{M}\mathcal{M}}$.

Results

Simulation studies

Simulation settings

In this section, we make a comprehensive comparison between FAGES and the traditional group variable selection methods: group LASSO[4], group MCP[5], and GEL[8]. The difference between FAGES and the traditional approaches is that the latter methods directly predict the response using the standard GLM (Eq. [3]), while FAGES considers the grouped factor-augmented GLM (Eq. [8]). We illustrate that FAGES outperforms the traditional approaches even though the underlying model is the standard GLM (Eq. [3]) rather than the grouped factor-augmented GLM (Eq. [8]).

We set $p_1 = \cdots = p_J = 100$ and $K_1 = \cdots = K_J = 3$. The sample size is set to be n = 200, 400, and 800, which reflects the cases of small, moderate, and large samples, respectively. The schemes to generate $\mathbf{X}$ and $\mathbf{F}$ are as follows. Each row of the factor matrix $\mathbf{F} = (\mathbf{F}_{\mathcal{M}_1}, \ldots, \mathbf{F}_{\mathcal{M}_J})$ is IID generated from $\mathcal{N}(0, \mathbf{R}_K(0.5))$, where $K = \sum_j K_j$ and $\mathbf{R}_K(r)$ denote the $(K \times K)$ AR(1) structure correlation matrix with correlation coefficient r. In each group $\mathcal{M}_j$, we generate $\mathbf{e}_{\mathcal{M}_j}$ from $\mathcal{N}(0, \mathbf{R}_{p_j}(0.5))$. To obtain the loadings matrix, we generate $\mathbf{s}_j = (s_j)_{p_j \times 1}$, where $s \sim \mathcal{U}(0,1)$, and calculate the $(p_j \times K_j)$ matrix of Tikhonov bases: $\boldsymbol{p}_{\mathcal{M}_j} = (\mathbf{1}, \boldsymbol{p}_1, \ldots, \boldsymbol{p}_{K_j-1})$, in which $\boldsymbol{p}_k = \sqrt{2}\cos(k\pi \boldsymbol{s})$. Denote $\mathbf{D}_{\mathcal{M}_j} = \mathrm{diag}(2\sqrt{2}, 2, \sqrt{2})$. Then, $\boldsymbol{\Lambda}_{\mathcal{M}_j} = \boldsymbol{p}_{\mathcal{M}_j}\mathbf{D}_{\mathcal{M}_j}$ and $\mathbf{X}_{\mathcal{M}_j} = \mathbf{F}_{\mathcal{M}_j}\boldsymbol{\Lambda}^{\top}_{\mathcal{M}_j} + \mathbf{e}_{\mathcal{M}_j}$. The group of non-zero coefficients is $\mathcal{M} = \bigcup_{j=1}^{J_0} \mathcal{M}_j$ with $J_0 = 10$, and the total number of groups is $J = 100$.

We consider the following two scenarios:

- Scenario 1. Assume the grouped factor-augmented GLM (Eq. [8]) holds: $\mathrm{E}(\mathbf{y}) = b'(\mathbf{F}\boldsymbol{\theta}^{\star})$. The coefficient $\boldsymbol{\theta}^{\star}_{\mathcal{M}_j}$ is randomly generated from $\mathcal{N}(0, \sigma^2 I_3)$ with $\sigma^2 = 9$. (The dimension of $\mathbf{F}_{\mathcal{M}_j}$ is 3 for all groups.)
- Scenario 2. Assume the standard GLM (3) holds: $\mathrm{E}(\mathbf{y}) = b'(\mathbf{X}\boldsymbol{\beta}^{\star})$. For each group, the first three entries of $\boldsymbol{\beta}^{\star}_{\mathcal{M}_j}$ are 1/12 and the last 97 entries are 0. (The dimension of $\mathbf{X}_{\mathcal{M}_j}$ is 100 for all groups.)

Scenario 1 is designed to simulate the real data, which have a factor structure in each group. FAGES is expected to outperform the traditional group variable selection approaches here. Scenario 2 is presented in order to investigate the robustness of FAGES. In this case, $\boldsymbol{\beta}^{\star}_{\mathcal{M}_j}$ is sparse and does not meet the setting of the grouped factor-augmented GLM. We conclude that FAGES is superior to the expectations if it still shows the best performance in this case. Note that we do not consider the intercept, i.e., setting $\theta_0 = 0$.

The evaluation criteria include the prediction error (PE), the computing time, the proportion of true positives (TP), and the proportion of true negatives (TN). Specifically, the PE is $\|\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}\|_2/\sqrt{n}$, where η is $\mathbf{F}\boldsymbol{\theta}^{\star}$ or $\mathbf{X}\boldsymbol{\beta}^{\star}$ corresponding to Scenario 1 or Scenario 2, respectively, and $\hat{\boldsymbol{\eta}}$ is $\hat{\mathbf{F}}\hat{\boldsymbol{\theta}}$ or $\mathbf{X}\hat{\boldsymbol{\beta}}$ depending on which approach we apply. The computing time in seconds is employed to verify the computational efficiency. Moreover, TP is the percentage of correctly estimated non-zero groups, and TN is the percentage of correctly estimated zero groups. The R package `grpreg`[28] is used to deal with the minimizations. The number of candidates of λ is taken as 100, and the Bayesian information criterion (BIC)[33] is used to determine the optimal λ. The number of replications of the simulations is 500 and the processor is 12th Gen Intel(R), Core(TM), i7-12700, 2.10 GHz, and16 GB.

To isolate and examine the intrinsic statistical properties of FAGES, we begin with the simulation settings in which the true number of latent factors is known and is treated as oracle information. This design allows us to disentangle the intrinsic performance of FAGES from the additional variability introduced by factor dimension estimation. In the Supplementary Material (Supplementary Fig. S1–S4), we further compare the results obtained using the oracle factor dimension with those based on the proposed adaptive selection strategy that combines multiple criteria. We find that, at least under these idealized simulation settings, the adaptive strategy can consistently recover the true number of factors, leading to a performance that is nearly indistinguishable from the one achieved using the oracle factor dimension.

Example 1: Linear regression

We first consider the linear regression model, i.e., $\mathbf{y} = \boldsymbol{\eta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^{\top}$ with $\epsilon_i \sim \mathcal{N}(0,1)$, and η is equal to $\mathbf{F}\boldsymbol{\theta}^{\star}$ or $\mathbf{X}\boldsymbol{\beta}^{\star}$ corresponding to Scenario 1 or Scenario 2, respectively. The results shown in Figs. 2 and 3 correspond to Scenario 1 and 2, respectively.

Our findings from Fig. 2 are as follows. In terms of the PE, FAGES performs uniformly better than the traditional methods, regardless of the group variable selection penalty applied. Besides, the traditional method with GEL is the only accurate method. Groups LASSO and grSCAD cannot select any group when the group size is very large. In terms of computing time, FAGES is computationally efficient than the traditional ones since it reduces the dimension of the GLM by the PCA before minimizing it. Regarding the TN, both the traditional methods and FAGES show high accuracy when concave penalties are used. For the TP, FAGES shows the top performance, followed by the traditional method with GEL, while the remaining two traditional methods show no power. In addition, FAGES cannot achieve the group selection consistency with the group LASSO penalty because this penalty lacks the oracle property[1].

In Fig. 3, we observe the following phenomena when the grouped factor-augmented GLM is misspecified. Traditional methods with GEL and FAGES are almost of the same accuracy in terms of PE. This phenomenon illustrates that FAGES is robust even when the underlying model is misspecified. As for TN and TP, the traditional approaches with GEL and FAGES are of the same accuracy. Regarding the computing time, FAGES is much faster than all three traditional methods, since it handles a model with a much lower dimension than the traditional methods. In summary, FAGES is still accurate when the underlying model is totally misspecified. In contrast, the traditional method performs well only when the GEL penalty is employed.

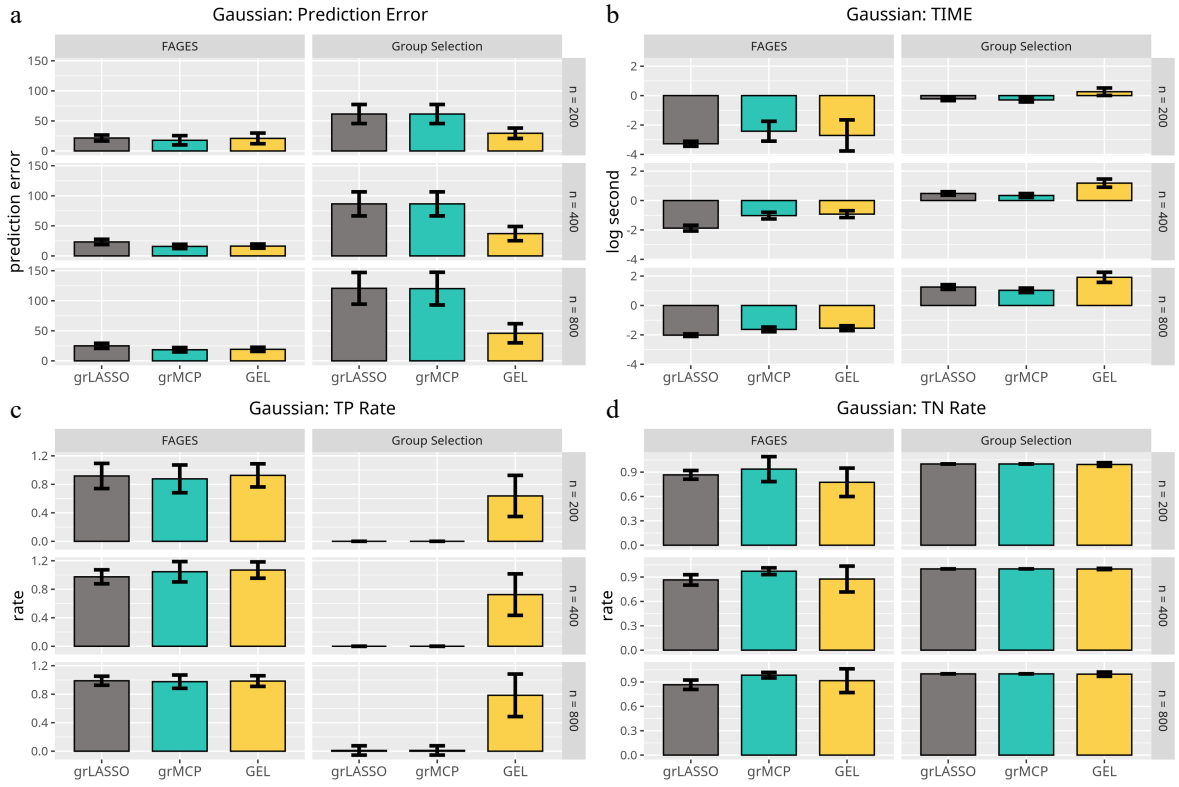

Fig. 2 Results of the linear model with respect to Scenario 1. Each panel represents a different evaluation metric: (a) prediction error, (b) computing time, (c) true negative rate, and (d) true positive rate. The height of each bar indicates the average value of the corresponding metric across multiple simulations, with error bars representing twice the standard deviation. The first three bars in each panel correspond to traditional methods, while the last three bars represent FAGES-based approaches with different penalty settings.

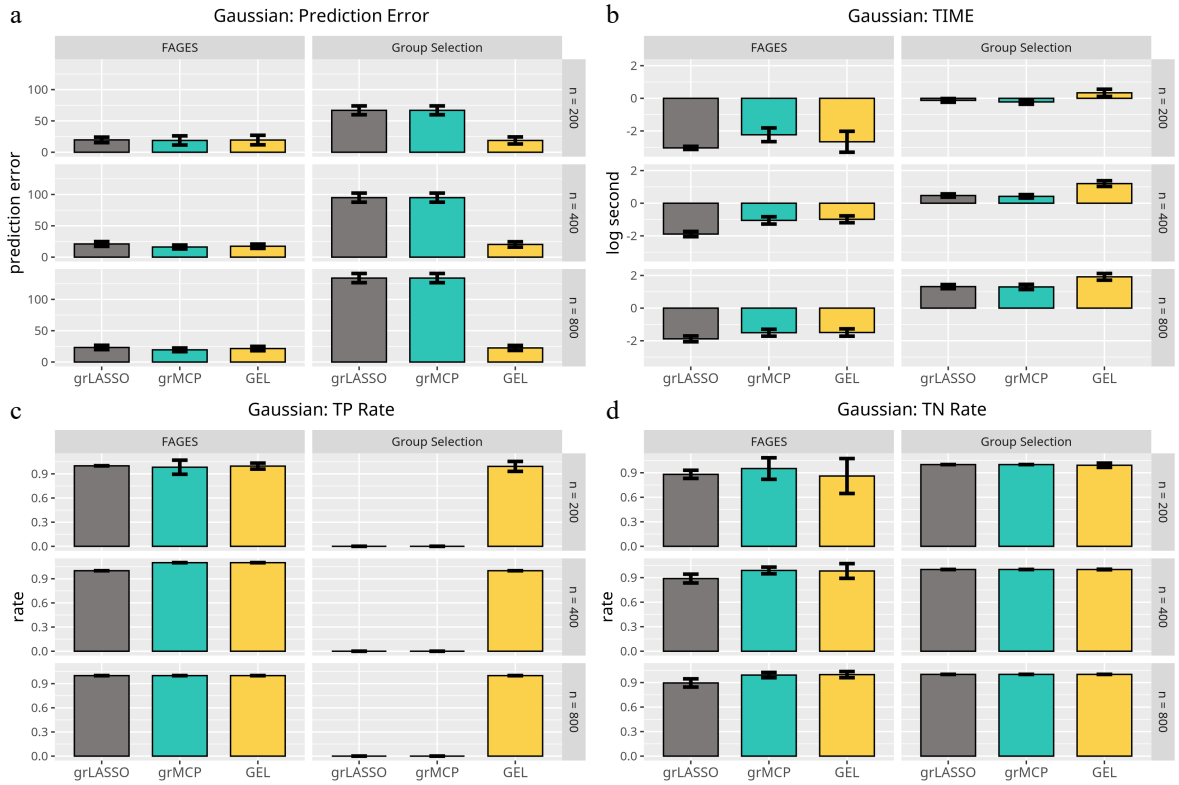

Fig. 3 Results of the linear model with respect to Scenario 2. Each panel represents a different evaluation metric: (a) prediction error, (b) computing time, (c) true positive rate, and (d) true negative rate. The height of each bar indicates the average value of the corresponding metric across multiple simulations, with error bars representing twice the standard deviation. The first three bars in each panel correspond to traditional methods, while the last three bars represent FAGES-based approaches with different penalty settings.

Example 2: Logistic regression

Now we investigate FAGES for binary responses under the logistic model: $\mathrm{E}(\mathbf{y}) = \exp(\boldsymbol{\eta})/\{1+\exp(\boldsymbol{\eta})\}$, where $\boldsymbol{\eta} = \mathbf{F}\boldsymbol{\theta}^{\star}$ in Scenario 1 and $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}^{\star}$ in Scenario 2. For binary outcomes, we observed that the group-penalized fitting in `grpreg` can be numerically unstable and may fail to converge, reflecting the fact that binary responses typically carry less information than Gaussian responses. To stabilize the optimization, we therefore use an elastic-net-type composite penalty by adding a ridge component with mixing parameter α in the minimization of θ. For the traditional approaches, we set $\alpha = 0.7$, which we found yields stable computation and avoids the extremely low TP rates that arise when the ridge penalty is omitted, in which case most groups are not selected. FAGES exhibits the same issue but to a lesser extent; accordingly, we use a slightly larger value $\alpha = 0.8$ to place less weight on the ridge penalty to reduce the bias. See `grpreg` for implementation details.

From Fig. 4, we learn that FAGES uniformly outperforms the traditional methods in terms of all four criteria when the underlying model is correctly specified. However, it should be noted that, despite the fact that traditional approaches perform far worse even when GEL is applied as a group selector, FAGES does not produce ideal results because it is likely to ignore groups with non-zero effects. Only when n is sufficiently large can FAGES achieve the model selection consistency. Fig. 5 illustrates that FAGES is more accurate than the traditional methods in terms of the PE, even when the underlying model is misspecified. The underlying explanation for this occurrence is that the binary response carries less information than the continuous response and thus, it prefers the model with less degrees of freedom, even if the other model with more degrees of freedom is closer to the true model. Hence, we conclude that FAGES is powerful and robust to analyze the high-dimensional grouped variables, especially for the binary response, as FAGES is able to greatly reduce the dimension of the conventional GLM (Eq. [3]) without losing any key information.

Example 3: Functional regression

We consider a more complex case: the observed variables of each group are in the form of smooth functions. Consider the following functional AFM:

$$\mathbf{X}(t) = \mathbf{F}\boldsymbol{\Lambda}(t)^{\top} + \mathbf{e}, \tag{18}$$

where, $t = (t_1, \ldots, t_p)^T$ is a covariate like the time of observation, $\lambda_1(t)$ is a smooth function of t, and $\boldsymbol{\Lambda}(t) = (\boldsymbol{\lambda}_1(t), \ldots, \boldsymbol{\lambda}_K(t))^{\top}$. In this case, the factor loading is considered a coordinate axis and the latent factor is viewed as the vector of random coordinates in this axis. To borrow information from the functional structure of $\boldsymbol{\Lambda}(t)$, Fan et al.[16] suggested first smoothing the observations $\mathbf{X}(t)$ by using a projection matrix $\mathbf{P}(t)$ and then yielding the consistent estimator of $\mathbf{F}$ and $\boldsymbol{\Lambda}(t)$ through the PCA. This method is known as the projection-PCA method. As for the choice of $\mathbf{P}(t)$, one may employ the Tikhonov bases if $t_1, \ldots, t_p$ are regularized in interval [0, 1]. Let $\boldsymbol{p}_s(t) = \sqrt{2}\cos(s\pi t)$ and $\boldsymbol{p}(t) = (\mathbf{1}, \boldsymbol{p}_1(t), \ldots, \boldsymbol{p}_S(t))$, where S is a given cutoff. Then $\mathbf{P}(t) = \boldsymbol{p}(t)\boldsymbol{p}^{\top}(t)/p$. For a suitable choice of S, for example, $S = 9$, it is guaranteed that $\boldsymbol{\Lambda}^{\top}(t)\mathbf{P}(t) \approx \boldsymbol{\Lambda}^{\top}(t)$ and $\mathbf{e}\mathbf{P}(t) \approx 0$[34]. As a result, analyzing the projected data $\hat{\mathbf{X}} = \mathbf{X}(t)\mathbf{P}(t)$ is an approximately noiseless problem, i.e.,

$$\hat{\mathbf{X}}\hat{\mathbf{X}}^{\top}/n \approx \mathbf{F}\{\boldsymbol{\Lambda}(t)^{\top}\mathbf{P}(t)\boldsymbol{\Lambda}(t)\}\mathbf{F}^{\top}/n. \tag{19}$$

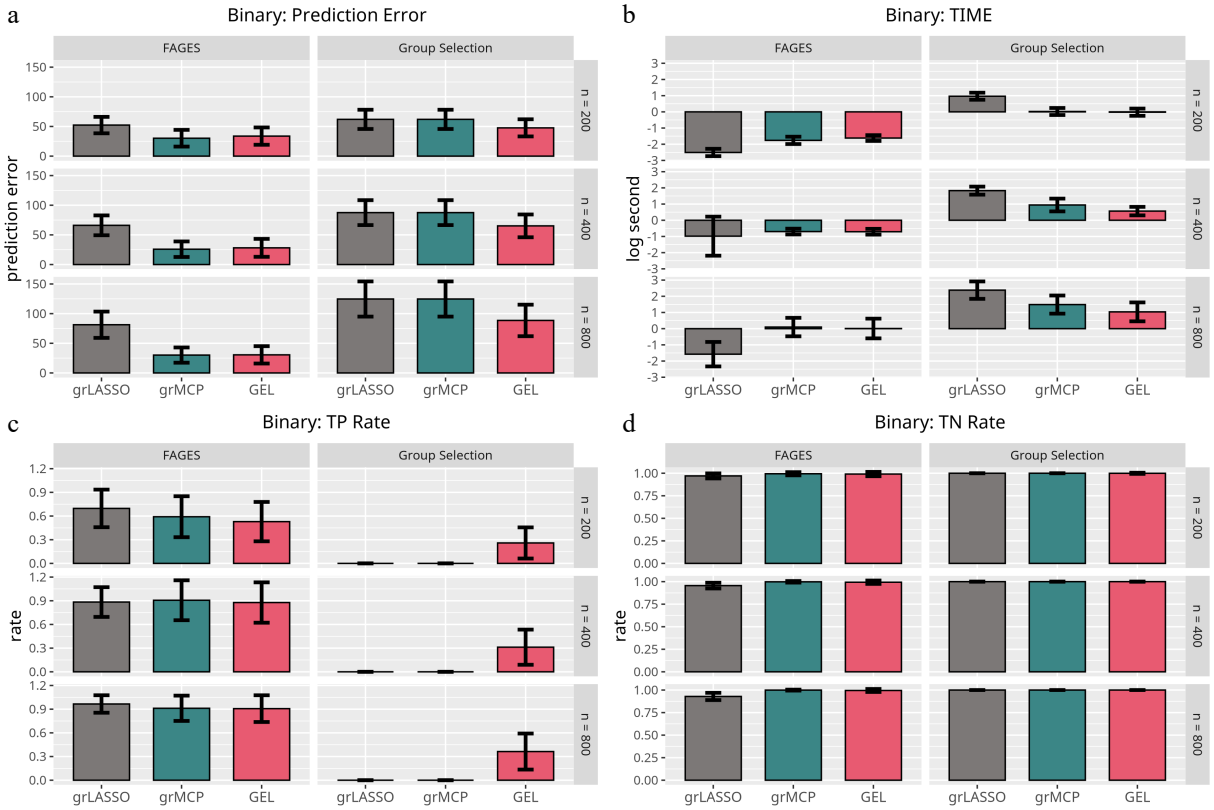

Fig. 4 Results of the logistic model with respect to Scenario 1. Each panel represents a different evaluation metric: (a) prediction error, (b) computing time, (c) true positive rate, and (d) true negative rate. The height of each bar indicates the average value of the corresponding metric across multiple simulations, with error bars representing twice the standard deviation. The first three bars in each panel correspond to traditional methods, while the last three bars represent FAGES-based approaches with different penalty settings.

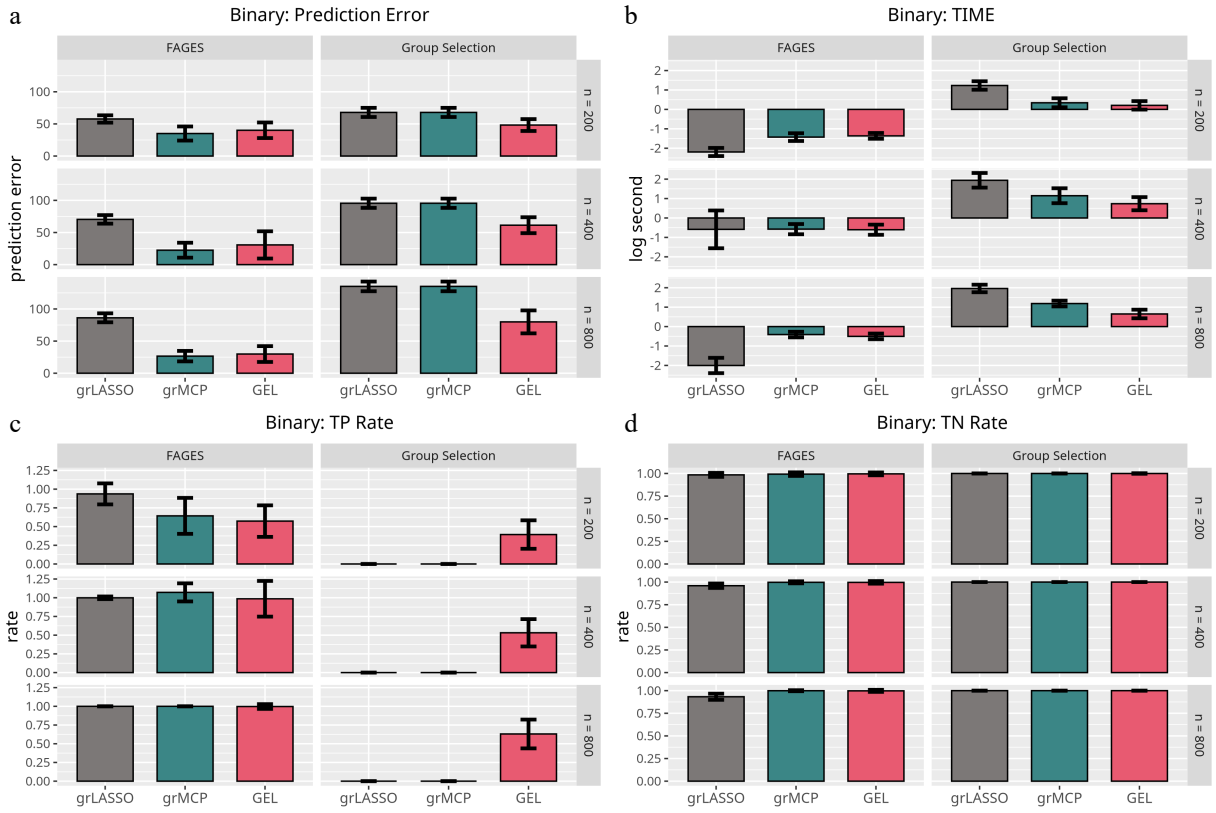

Fig. 5 Results of the logistic model with respect to Scenario 2. Each panel represents a different evaluation metric: (a) prediction error, (b) computing time, (c) true negative rate, and (d) true positive rate. The height of each bar indicates the average value of the corresponding metric across multiple simulations, with error bars representing twice the standard deviation. The first three bars in each panel correspond to traditional methods, while the last three bars represent FAGES-based approaches with different penalty settings.

Similar to the ordinal AFM, the first K eigenvectors of $\hat{\mathbf{X}}\hat{\mathbf{X}}^\top/n$ form the estimator of $\hat{\mathbf{F}}/\sqrt{n}$, and $\hat{\Lambda}(t) = \hat{\mathbf{X}}^\top\hat{\mathbf{F}}/\sqrt{n}$ is the latent loading estimator.

Without losing generality, we set $p_1 = \cdots = p_J = 2{,}000$, and other parameters remain the same as in the previous simulations. Although the number of real observation points of a curve may be larger than 20,000, our experience shows it is sufficient to accurately estimate the function by using 2,000 points randomly sampled from the 20,000. We apply the projection-PCA to smooth the sample matrix $\mathbf{X}_{\mathcal{M}_j}(t)$ by using the projection matrix $\mathbf{P}(t) = \boldsymbol{p}(t)\boldsymbol{p}^\top(t)/p$, where $\boldsymbol{p}(t) = (\mathbf{1}, t_1(t), \ldots, t_9(t))$ is the matrix of the Tikhonov basis functions. The latent factor $\hat{\mathbf{F}}_{\mathcal{M}_j}$ is estimated from the smoothed matrix $\hat{\mathbf{X}}_{\mathcal{M}_j} = \mathbf{X}_{\mathcal{M}_j}(t)\mathbf{P}(t)$ through the minimization (Eq. [2]). Here, we do not compare FAGES with the traditional group variable selection methods since the observation points of each function may be very large, in which case the traditional methods are very likely to fail.

The top three rows of Fig. 6 show the results of the linear model for functional data analysis. FAGES is clearly capable of handling the data that are in the form of smooth curves. FAGES accomplishes the model selection consistency with a high probability using the concave penalty no matter the sample size is large or small. The bottom three rows of Fig. 6 present the counterpart of the logistic regression model. FAGES is likely to ignore certain non-zero groups, consistent with the previous simulations. Only when the sample size is sufficiently large can FAGES achieve the model selection consistency.

Real-data analysis

Data

In this section, we demonstrate the analysis of the HNSCC data through FAGES. We first give a description of the data and related concepts and then show the results of the analysis. The HNSCC data are provided by the cancer genome atlas (TCGA) program and can be found on the website UCSC Xena (https://xenabrowser.net/datapages). The website contains data related to n = 520 patients and 20,530 genes, which show the gene-level transcription estimates, as in the $\log_2(x+1)$-transformed RSEM normalized count. The gene "Ki67" is considered the response variable because it is the indicator gene in many biological researches. For example, "Ki67" may be necessary for cellular proliferation; it is involved in maintaining the individual mitotic chromosomes dispersed in the cytoplasm following nuclear envelope disassembly; and higher expression of "Ki67" is related to a poor prognosis of cancer[35,36].

A pathway is a kind of gene set commonly used in biological researches. There are more than 20,000 genes in humans, and their molecular functions are based on the so-called biological pathways, which host a series of interactions among genes or molecules in a cell that lead to a biological function. KEGG is a kind of pathway database built for understanding the high-level functions and utilities of the biological system from gene-level or molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies[37].

The KS random walk method[9] is applied to convert the gene expressions into an individualized KS curve that reflects the

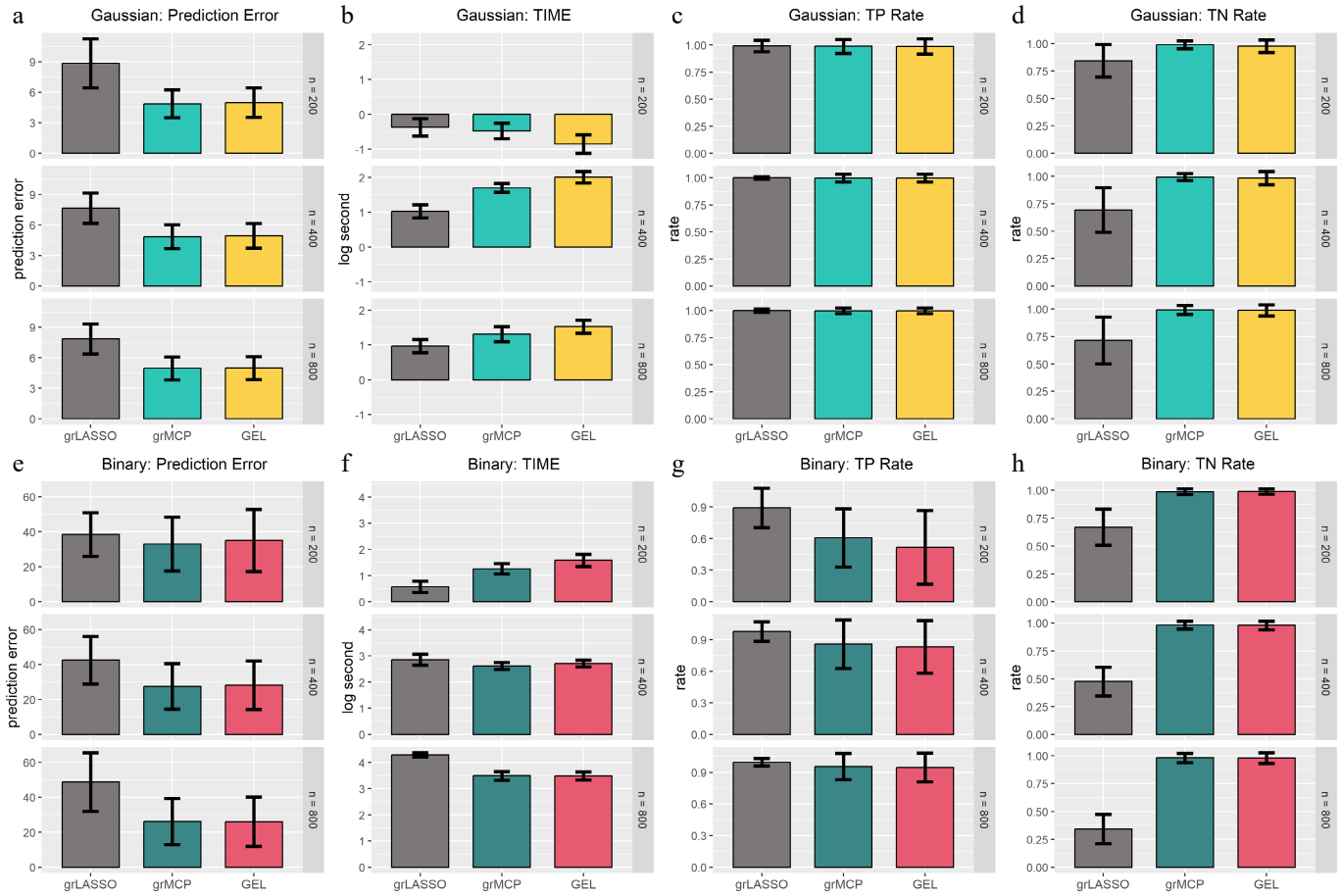

Fig. 6 Results of functional data analysis. The first four rows correspond to the Gaussian model, while the last four rows correspond to the binary model. Each column represents a different evaluation metric: (a), (e) prediction error; (b), (f) computing time; (c), (g) true negative rate; and (d), (h) true positive rate. The height of each bar indicates the average value of the corresponding metric across multiple simulations, with error bars representing twice the standard deviation. Different methods are compared, with the first set of bars corresponding to traditional approaches and the latter set representing FAGES-based methods with different penalty settings.

enrichment of a target gene set. Suppose there is a matrix of expression profiles $\mathbf{V} = (V_{ij})_{n\times p}$, where n is the number of samples and p is the number of genes. The KS random walk method first ranks the V_{ij} for every sample to obtain the rank statistics S_{ij} and then yields the KS curve using the accumulation process

$$\boldsymbol{X}_{il,\mathcal{M}_j} = \frac{\sum_{h=1}^{l} |S_{ih}| \mathbf{1}(h \in \mathcal{M}_j)}{\sum_{h=1}^{p} |S_{ih}| \mathbf{1}(h \in \mathcal{M}_j)} - \frac{\sum_{h=1}^{l} \mathbf{1}(h \notin \mathcal{M}_j)}{p - |\mathcal{M}_j|}, \tag{20}$$

where, $\mathbf{1}(h \in \mathcal{M}_j)$ is the indicator function that shows whether the h-th gene (the gene corresponding to the h-th ranked expression-level statistic) belongs to gene set $\mathcal{M}_j$, $|\mathcal{M}_j|$ is the number of genes in the j-th gene set, and p is the number of genes in the gene data. For $l = 1, \ldots, p$, the KS statistics $\boldsymbol{X}_{il,\mathcal{M}_j}$ form an individualized KS curve corresponding to gene set $\mathcal{M}_j$.

Analysis

We consider the KEGG pathways as gene sets and utilize the KS random walk approach to generate the individualized KS curves for every pathway. Besides, we assume the individualized KS curves to follow an AFM (Eq. [7]) in each pathway, and use the projection-PCA to estimate the related latent factor. Without loss of generality, the dimensions of the latent factors are uniformly set as 3 and only the 1st,..., 1027th quantiles of each KS curve are recorded. We use the penalized likelihood (Eq. [15]) to handle the grouped factor-augmented GLM (Eq. [8]). The averaged group effect is computed through Eq. (16) and the effect direction is determined based on Eq. (17).

Figure 7 demonstrates the identified pathways and the related pathway effects using FAGES. First of all, FAGES with the two different penalties identifies many consistent pathways, including Toxoplasmosis, TGF-beta, *Staphylococcus aureus* infection, Progesterone edited oocyte maturation, etc. Besides, FAGES with the GEL identifies ten more important pathways than FAGES with the grMCP because of GEL's flexibility. The cell cycle and TGF-beta are the two pathways that have the most significant averaged group effects in both methods. The cell cycle pathway, which regulates DNA synthesis and mitosis during tumor cell proliferation, is recognized to be positively associated with tumor cell proliferation[38]. The TGF-beta pathway, consistent with previous reports, has a strongly negatively averaged group effect on cell proliferation, implying that the activation of this pathway might compromise tumor proliferation[39]. Furthermore, pathways with considerable averaged group effects, such as prostate cancer, toxoplasmosis, amino sugar and nucleotide sugar metabolism, and inositol phosphate metabolism, warrant further investigation. In this process, FAGES acts as a screen that removes the noisy genes in a more statistically valid and biologically understandable manner.

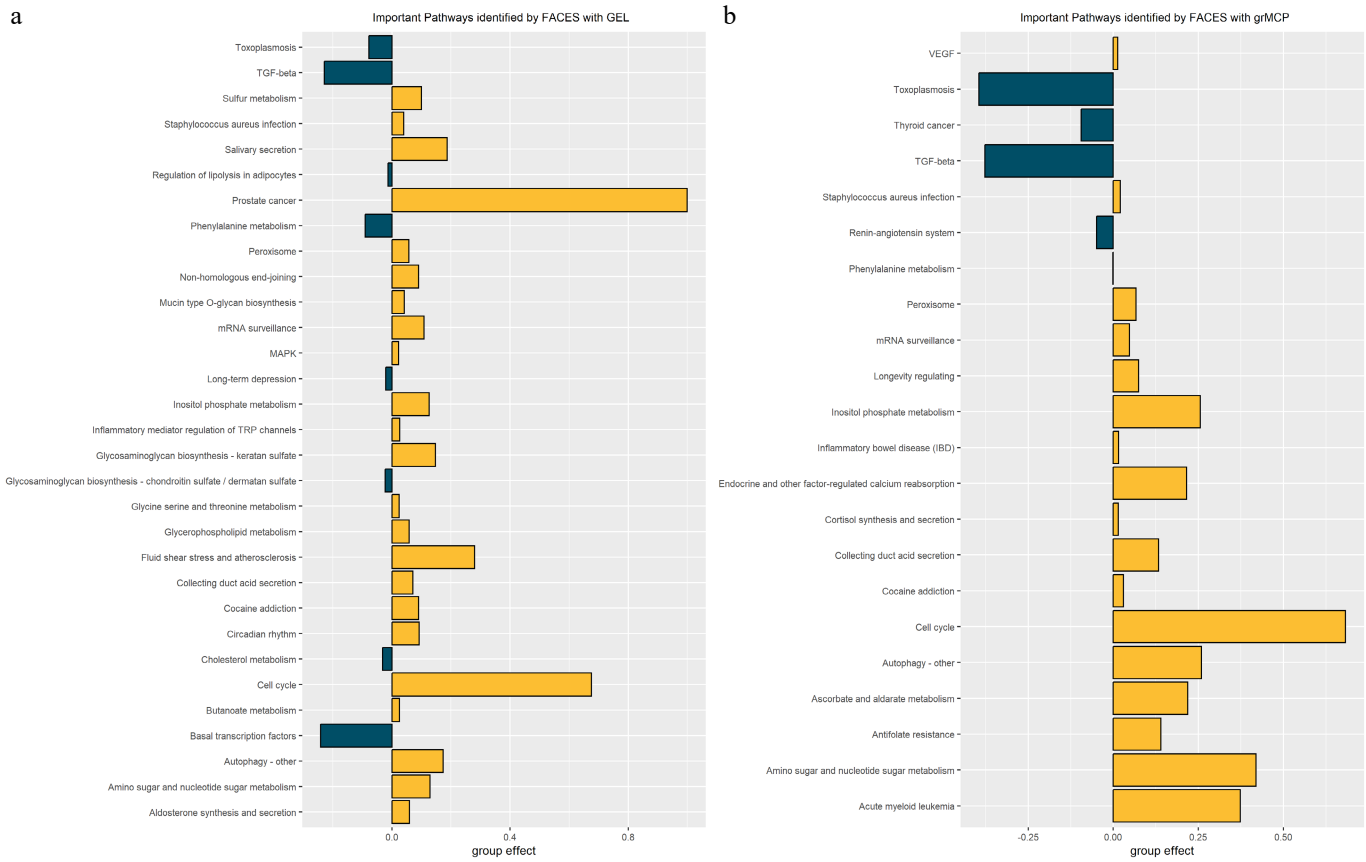

Fig. 7 Important pathways identified by FAGES with different penalty functions. (a) shows the results obtained using GEL, while (b) displays the results yielded by grMCP. The bars represent the estimated averaged group effect for each selected pathway, with the pathway names shown on the y-axis. Green bars indicate pathways that are positively associated with the outcome, while yellow bars indicate pathways that are negatively associated with the outcome.

Discussion

In this paper, we introduce a two-stage method called FAGES for identifying important variable groups within the framework of GLM. Methodologically, FAGES first addresses within-group correlations by modeling each group of variables with an AFM. Each group is assumed to be driven by a few latent factors together with idiosyncratic components, with the latent factors capturing the dominant variation[18]. By representing groups through their latent factors, FAGES reduces dimensionality and specifies a grouped factor-augmented GLM in which these latent factors are entered as predictors associated with the outcome. A variable selection penalty is then applied to simultaneously identify relevant groups and estimate their corresponding regression coefficients. For inference, FAGES quantifies the strength of the selected groups through an averaged group effect estimate (Eq. [16]), which provides a comparable measure of the group contribution.

In our simulations, we found that grMCP performs slightly better than GEL. This is not surprising. Under idealized settings, grMCP does not impose an explicit bi-level selection structure and can therefore exploit the group information more efficiently, which may translate into higher power. However, prior empirical evidence in the literature suggests that, in real-data applications, GEL tends to be more robust to heterogeneous within-group correlation patterns and model misspecification[8]. Hence, we recommend GEL as the default choice in practice, and view grMCP as a competitive alternative when the signal is sufficiently strong and the group structure is well aligned with the assumed model.

As an application of FAGES, we show how to enhance GSVA by using FAGES in real-data analysis. Traditional GSVA relies on a single summary statistic, such as the maximum deviation or maximum difference statistic, to extract information from the KS curve[9]. We propose leveraging the leading FPCs of the KS curve to more effectively capture the contribution of gene sets to the outcome, and estimate the direct effect of a gene set on the outcome conditional on other gene sets. Beyond this application, we further hypothesize that FAGES can benefit Mendelian randomization (MR) analyses based on rare variants[40,41]. Specifically, we propose to ① construct a weighted kernel matrix for rare variants within a genomic region based on SKAT[42], and ② apply PCA to extract the leading kernel-weighted PCs. Furthermore, we can ③ select key genomic regions and estimate the effects of their PCs on both the exposure and the outcome, which can serve as instrument variables (IVs) in MR, and ④ ultimately perform MR to assess whether the exposure causally influences the outcome through the rare variants. For example, rare loss-of-function variants in ANGPTL3 have been associated with a decreased risk of cardiovascular diseases[43], while the common *cis* protein quantitative trait locus (pQTL) of ANGPTL3 abundance showed no causality[44]. In the future, we can investigate whether lipid traits, such as triglyceride, mediate their effect on ASCVD throughrareloss-of-functionANGPTL3variantsbyusingaFAGE-basedMR analysis.

Several limitations of the present study should be acknowledged. First, in the real-data analysis, we chose Ki67 expression as the response variable. Although Ki67 is a well-established marker of

cellular proliferation and is widely recognized as a prognostic indicator in oncology[35,36], modeling the expression level of a single gene is less conventional in GSA. Thus, this application should be viewed mainly as a proof-of-concept illustration of FAGES. Second, the two-stage estimation procedure, which first extracts latent factors and then applies penalized regression, may introduce additional variability and potential heteroskedasticity[45]. Moreover, we did not include direct comparisons with classical GSA methods (e.g., AUCell[46] and PAGODA[47]), because these approaches are marginal by design, while our aim is to advance group variable selection methods that estimate the conditional effects in a regression framework.

Author contributions

The author confirms sole responsibility for all aspects of this study and approved the final version of the manuscript.

Data availability

All data generated or analyzed during this study are included in this published article and its supplementary materials.

Acknowledgments

We would like to thank the editors and the anonymous referees for their very careful reviews and suggestions.The first author would like to thank Dr. Jianxin Pan and Dr. Hao Xu for their multiple inspiring discussions.

Conflict of interest

The author declares that there is no conflict of interest.

Supplementary information accompanies this paper online at: https://doi.org/10.48130/stati-0026-0007.

Dates

Received 30 January 2026; Revised 31 January 2026; Accepted 19 March 2026; Published online 29 May 2026

References

[1] Fan J, Li R. 2001. Variable selection *via* nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96(456):1348–1360

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

[3] Zhang CH. 2010. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* 38(2):894–942

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

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

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

[7] Huang J, Ma S, Xie H, Zhang CH. 2009. A group bridge approach for variable selection. *Biometrika* 96(2):339–355

[8] Breheny P. 2015. The group exponential lasso for bi-level variable selection. *Biometrics* 71(3):731–740

[9] Hänzelmann S, Castelo R, Guinney J. 2013. GSVA gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* 14(1):7

[10] Kanehisa M. 2000. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* 28(1):27–30

[11] Chandler G, Polonik W. 2021. Multiscale geometric feature extraction for high-dimensional and non-euclidean data with applications. *The Annals of Statistics* 49(2):988–1010

[12] Bach FR. 2008. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research* 9(6):1179–1225

[13] Zhao P, Yu B. 2006. On model selection consistency of lasso. *Journal of Machine Learning Research* 7:2541–2563

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

[15] Fan J, Ke Y, Wang K. 2018. Factor-adjusted regularized model selection. *SSRN Electronic Journal* 216(1):71–85

[16] Fan J, Liao Y, Wang W. 2016. Projected principal component analysis in factor models. *The Annals of Statistics* 44(1):219–254

[17] Yao F, Müller HG, Wang JL. 2005. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* 100(470):577–590

[18] Bai J. 2003. Inferential theory for factor models of large dimensions. *Econometrica* 71(1):135–171

[19] Ramsay JO, Silverman BW. 2005. *Functional data analysis*. New York, US: Springer Science & Business Media. doi: 10.1007/b98888

[20] Nelder JA, Wedderburn RWM. 1972. Generalized linear models. *Journal of the Royal Statistical Society Series A (General)* 135(3):370–384

[21] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. 2000. Gene Ontology: tool for the unification of biology. *Nature Genetics* 25(1):25–29

[22] Fan J, Lv J. 2011. Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory* 57(8):5467–5484

[23] Bai J, Ng S. 2006. Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica* 74(4):1133–1150

[24] Hall P, Müller HG, Yao F. 2008. Modelling sparse generalized longitudinal observations with latent Gaussian processes. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 70(4):703–723

[25] Onatski A. 2009. Testing hypotheses about the number of factors in large factor models. *Econometrica* 77(5):1447–1479

[26] Lam C, Yao Q. 2012. Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics* 40(2):694–726

[27] Fan J, Guo J, Zheng S. 2022. Estimating number of factors by adjusted eigenvalues thresholding. *Journal of the American Statistical Association* 117(538):852–861

[28] Breheny P, Zeng Y, Kurth R, Breheny MP. 2024. *Package grpreg*. https://cran.r-project.org/web/packages/ncvreg/index.html

[29] Li G, Peng H, Zhang J, Zhu L. 2012. Robust rank correlation based screening. *The Annals of Statistics* 40(3):1846–1877

[30] Fan J, Liao Y, Mincheva M. 2013. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 75(4):603–680

[31] Wainwright MJ. 2019. *High-dimensional statistics: a non-asymptotic viewpoint*. Vol. 48. UK: Cambridge University Press. doi: 10.1017/9781108627771

[32] Fan J, Xue L, Zou H. 2014. Strong oracle optimality of folded concave penalized estimation. *The Annals of Statistics* 42(3):819–849

[33] Schwarz G. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6(2):461–464

[34] Wood SN. 2017. *Generalized additive models: an introduction with R*. New York, US: CRC Press. doi: 10.1201/9781315370279

[35] Cuylen S, Blaukopf C, Politi AZ, Müller-Reichert T, Neumann B, et al. 2016. Ki-67 acts as a biological surfactant to disperse mitotic chromosomes. *Nature* 535(7611):308–312

[36] Scholzen T, Gerdes J. 2000. The Ki-67 protein: from the known and the unknown. *Journal of Cellular Physiology* 182(3):311–322

[37] Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research* 45(D1):D353–D361

[38] Evan GI, Vousden KH. 2001. Proliferation, cell cycle and apoptosis in cancer. *Nature* 411(6835):342–348
[39] Huang SS, Huang JS. 2005. TGF-β control of cell proliferation. *Journal of Cellular Biochemistry* 96(3):447–462
[40] Lorincz-Comi N, Yang Y, Li G, Zhu X. 2024. MRBEE: a bias-corrected multivariable Mendelian randomization method. *Human Genetics and Genomics Advances* 5(3):100290
[41] Sanderson E, Glymour MM, Holmes MV, Kang H, Morrison J, et al. 2022. Mendelian randomization. *Nature Reviews Methods Primers* 2:6
[42] Wu MC, Lee S, Cai T, Li Y, Boehnke M, et al. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* 89(1):82–93
[43] Dewey FE, Gusarova V, Dunbar RL, O'Dushlaine C, Schurmann C, et al. 2017. Genetic and pharmacologic inactivation of ANGPTL3 and cardiovascular disease. *The New England Journal of Medicine* 377(3):211–221
[44] Landfors F, Henneman P, Chorell E, Nilsson SK, Kersten S. 2024. Drug-target Mendelian randomization analysis supports lowering plasma ANGPTL3, ANGPTL4, and APOC3 levels as strategies for reducing cardiovascular disease risk. *European Heart Journal Open* 4(3):oeae035
[45] Zhang AR, Cai TT, Wu Y. 2022. Heteroskedastic PCA: algorithm, optimality, and applications. *The Annals of Statistics* 50(1):53–80
[46] Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, et al. 2017. SCENIC: single-cell regulatory network inference and clustering. *Nature Methods* 14(11):1083–1086
[47] Fan J, Salathia N, Liu R, Kaeser GE, Yung YC, et al. 2016. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nature Methods* 13(3):241–244