Feature grouping and sparse principal component analysis with truncated regularization

Haiyan Jiang1 | Shanshan Qin2 | Oscar Hernan Madrid Padilla3

1School of Statistics and Data Science, Nankai University, Tianjin, China
2School of Statistics, Tianjin University of Finance and Economics, Tianjin, China
3Department of Statistics, University of California, Los Angeles, Los Angeles, California, USA

Correspondence
Shanshan Qin, School of Statistics, Tianjin University of Finance and Economics, Tianjin, China.
Email: qinsslzu@gmail.com

Funding information
Tianyuan Fund for Mathematics, National Natural Science Foundation of China, Grant/Award Number: 12226333; National Natural Science Foundation of China, Grant/Award Number: 12201454

We propose a new method for principal component analysis (PCA) called feature grouping and sparse principal component analysis (FGSPCA). This method is designed to capture both grouping and sparsity structures in factor loadings simultaneously. To achieve this, we use a non-convex truncated regularization that can adjust for sparsity and grouping effects automatically. This regularization encourages factor loadings with similar values to be either grouped together for feature grouping or be zero for feature selection, helping to reduce model complexity and improve interpretation. While other structured PCA methods often require prior knowledge to construct the regularization term, FGSPCA can capture grouping and sparsity structures without any prior information. We solve the resulting non-convex optimization problem using an alternating algorithm that combines difference-of-convex programming, the augmented Lagrange method, and coordinate descent method. Our experiments show that FGSPCA performs well and efficiently on both synthetic and real-world datasets. An implementation of FGSPCA is available on GitHub https://github.com/HaiyanJiang/FGSPCA.

KEYWORDS
feature grouping, feature selection, non-convex truncated regularization, principal component analysis, sparsity

INTRODUCTION

Principal component analysis (PCA) (Jolliffe, 1986) is an important unsupervised technique for feature extraction and dimension reduction, with numerous applications in statistics and machine learning, such as gene representation and face recognition. The goal of PCA is to find a sequence of linear combinations of the original variables/predictors by projecting the original data onto an orthogonal linear space, called principal components (PCs), such that the derived PCs capture the maximum variance along the orthogonal direction. Numerically, PCA can be obtained via the singular value decomposition (SVD) of the data matrix. Denote $X_{n \times p} \in \mathbb{R}^{n \times p}$ a data matrix consisting of $n$ observations of a random vector $x \in \mathbb{R}^p$ with a population covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$, where $n$ and $p$ are the number of observations and the number of variables/predictors, respectively. Without loss of generality, assume that all the predictors are centred with zero means. Let the SVD of $X$ be $X = UDV^T$. The projection of the data $Z = UD(= XV)$ are the derived PCs, and the columns of $V$ are the corresponding factor loadings (or factor coefficients, or PC vectors).

PCA aims to recover the top $k$ leading eigenvectors $u_1, \ldots, u_k$ of the population covariance matrix $\Sigma$, with the corresponding eigenvalues $\lambda_1 \geq \ldots \geq \lambda_k$. In high dimensional settings with $p \gg n$, the ordinary PCA can be inconsistent (Johnstone & Lu, 2009; Nadler, 2008; Paul, 2007), and additional assumptions are needed to avoid the curse of dimensionality (Wang et al., 2014). Besides, a simple property of the ordinary PCA is that each PC usually involves all the original variables and the loadings (factor coefficients) are typically nonzero, which hinders the interpretability of the derived PCs. In order to deal with the curse of dimensionality and improve the interpretability of the derived PCs, a sparsity assumption is often imposed on the loadings to get a sparsely weighted linear combination of the original variables. PCA with sparse loadings and its variants (Cai et al., 2013; Erichson et al., 2020; Vu et al., 2013; Zou et al., 2006) have been widely studied. In the last decades, significant progress has been
made on the methodological development as well as theoretical understanding of sparse PCA. One can turn to Erichson et al. (2020), Wang et al. (2014), Jenatton et al. (2011), Jenatton et al. (2010), Grbovic et al. (2012), Croux et al. (2013), Khan et al. (2015), Yi et al. (2017), Zou and Xue (2018), Jin and Sidford (2019), Zhang and Tong (2019), and Tian et al. (2020), among others, for an overview of the literature. Methods introduced in these articles intend to seek modified principal components for various sparsity properties. For example, SCoTLASS (Jolliffe et al., 2003) is proposed by directly imposing an \( \ell_1 \) penalty on the ordinary PC vectors to get sparse loadings. Sparse PCA (SPCA) (Zou et al., 2006) seeks sparse loadings by extending the elastic net (Zou & Hastie, 2005) procedure and relaxing the orthogonality constraint of the ordinary PC vectors.

In addition to the sparsity property among loadings, the structured grouping property can also lead to good interpretability of the resulting PCs. SPCA (Zou et al., 2006) can achieve better interpretability by producing modified PCs with sparse loadings. However, it does not take into account the structured grouping property among loadings, that is, clusters or groups. Based on the structured variable selection method (Jenatton et al., 2011), a structured sparse PCA (Jenatton et al., 2010) is proposed to explore the structural information, as an extension of sparse PCA, and it incorporates prior knowledge into the sparsity-inducing regularization and is able to encode more sophisticated sparsity patterns. In order to capture the “blocking” structures in the factor loadings, Guo et al. (2010) proposed another variant of PCA with sparse fused loadings, named sparse fused PCA (SFPCA), by introducing a fusion penalty that encourages the loadings associated with high correlation to be close to get the “blocking” structures. Recently, Tian et al. (2020) proposed the feature-sparsity (row-sparsity) constrained PCA by considering feature-sparsity structures for feature selection and PCA simultaneously. However, these methods depend heavily on the structured prior knowledge, which is usually challenging to obtain or specify in real applications. In Guo et al. (2010), for example, the “blocking” structure is captured by the fusion penalty, where the fusion penalty depends on the sample correlation which serves as the prior information. Moreover, even though the PC vectors derived from the structured sparse PCA possess some sparse structures, they suffer from the same issue; that is, the structured sparsity depends on the given structural prior information.

In the ordinary PCA, each PC is a linear combination of all \( p \) variables, and the loadings are typically nonzero and have no grouping effect. As is discussed above, the loadings can be sparse in sparse PCA (Zou et al., 2006), but dismissing grouping effect or clustering effect among the loadings. In structured PCA (Guo et al., 2010; Jenatton et al., 2010; Tian et al., 2020), the structures of the loadings can be learned based on the structural prior knowledge/information, which should be given to construct the regularization term in these methods.

In this paper, we propose a new variant of PCA, named feature grouping and sparse principal component analysis (FGSPCA), which can simultaneously capture the grouping and sparse structure of factor loadings, leading to modified PCs with grouping- and sparse-guided loadings. By adopting the fact that PCA can be formulated as a regression-type optimization problem, the grouping- and sparse-guided loadings are obtained by imposing the grouping and sparsity constraints on the regression coefficients. We make the following contributions.

- To our knowledge, we are among the first to consider simultaneously the grouping effect as well as the sparsity effect among factor loadings of PCA in the absence of prior knowledge. The proposed FGSPCA method achieves the goal of feature grouping and feature selection via regularization, whose construction does not depend on any prior knowledge. The grouping and sparsity structure is learned naturally from the model rather than from given prior information.
- The proposed FGSPCA method imposes a non-convex regularization term with naturally adjustable sparsity and grouping effect. We solve the non-convex FGSPCA problem approximated by a sequence of linear convex subproblems via the difference-of-convex programming (DC). Each of the convex subproblems is solved iteratively by incorporating the augmented Lagrange method (AL) and the coordinate descent method (CD).
- The experiments on both synthetic and real-world data demonstrate the promising performance of the proposed FGSPCA method.

Throughout this paper, we use the following notations. Boldface lowercase letters refer to vectors, for example, \( \mathbf{a}, \mathbf{b} \), and boldface uppercase letters refer to matrices, for example, \( \mathbf{A}, \mathbf{B} \). For a vector \( \mathbf{w} \in \mathbb{R}^p \), denote by \( \| \mathbf{w} \|_2^2 = \sum_{j=1}^{p} w_j^2, \| \mathbf{w} \|_1 = \sum_{j=1}^{p} |w_j| \), and \( \| \mathbf{w} \|_\infty = \max_{j \in \{1, \ldots, p\}} \{ |w_j| \} \) the squared \( \ell_2 \) norm, the \( \ell_1 \) norm, and the maximum norm, respectively. For a matrix \( \mathbf{W} \), let \( \| \mathbf{W} \|_F = \sqrt{\sum_{i,j} w_{ij}^2} \) denote the Frobenius norm. Note that \( \| \mathbf{W} \|_F^2 = \text{tr}(\mathbf{W}^T \mathbf{W}) \). Denote by \( \mathbf{I}_{k \times k} \), the identity matrix in \( \mathbb{R}^{k \times k} \). Let \( \mathbf{I}_{|A|} = 1 \) if the condition \( A \) holds; otherwise, \( \mathbf{I}_{|A|} = 0 \).

The rest of the paper is organized as follows. In Section 2, the PCA is revisited. Section 3 introduces the proposed FGSPCA and its connections to other sparse PCA variants. We propose an alternating algorithm to solve the FGSPCA problem in Section 4. Experiments to show the performance of FGSPCA and comparisons with other dimension reduction methods are presented in Section 5. A discussion on the extension of FGSPCA to the settings with non-negative loadings falls into Section 6. We conclude the paper in Section 7.

## 2 PRINCIPAL COMPONENT ANALYSIS REVISITED

Let \( \mathbf{X} = (x_{ij})_{n \times p} \) denote a data matrix with \( n \) observations and \( p \) variables. Assume that the columns of \( \mathbf{X} \) are all centred. In PCA, each PC is obtained by constructing linear combinations of the original variables that maximize the variance. Denote the SVD of \( \mathbf{X} \) by \( \mathbf{X} = \mathbf{U} \mathbf{D} \mathbf{V}^T \). Let \( \mathbf{Z}_t = \)
\textbf{U}D_j be the \(j\)-th PC, and columns of \(V = [V_1, ..., V_k]\) be the PC vectors or PC loadings. Except for the SVD decomposition, another way to derive the PC vectors is to solve the following constrained least squares problem,

\[
\min_A \|X - XAA^T\|_F^2, \quad \text{s.t. } A^TA = I_{k \times k},
\]

where \(A = [a_1, ..., a_k] \) is a \(p \times k\) matrix with orthogonal columns. The estimated \(A\) contains the first \(k\) PC vectors, and the projection of the data \(Z = XA\) is the first \(k\) PCs.

By relaxing the orthogonality requirement and imposing an \(\varepsilon_2\) penalty, Zou et al. (2006) proposed and reformulated PCA as the following regularized regression optimization problem, which is defined in Lemma 1 (Theorem 3 in Zou et al., 2006), see Appendix A in Supporting Information.

\textbf{Lemma 1.} Consider the first \(k\) principal components. Let \(x_i\) be the \(i\)-th row of data matrix \(X\). Denote \(A_{p \times k} = [a_1, ..., a_k], B_{p \times k} = [\beta_1, ..., \beta_k].\)

For any \(\lambda > 0\), let

\[
(\hat{A}, \hat{B}) = \arg \min_{A, B} \sum_{i=1}^n \|x_i - AB^T x_i\|_F^2 + \lambda \sum_{j=1}^k \|\beta_j\|_2^2, \quad \text{s.t. } A^TA = I_{k \times k}.
\]

Then \(\hat{\beta}_j \propto V_j\) for \(j = 1, ..., k\).

The PCA problem is transformed into a regression-type optimization problem with orthonormal constraints on \(A\), and all the sequences of principal components can be derived through Lemma 1. With the restriction \(B = A\), the objective function becomes \(\sum_{i=1}^n \|x_i - AB^T x_i\|_F^2 = \sum_{i=1}^n \|x_i - AA^T x_i\|_F^2\), whose minimizer under the orthonormal constraint on \(A\) consists exactly of the first \(k\) PC vectors of the ordinary PCA. Lemma 1 shows that the exact PCA can still be obtained by relaxing the restriction \(B = A\) and adding the ridge penalty term.

Note that

\[
\sum_{i=1}^n \|x_i - AB^T x_i\|_F^2 = \|X - XBA^T\|_F^2.
\]

Since \(A\) is orthonormal, let \(A_\perp\) be any orthonormal matrix such that \([A; A_\perp]\) is \(p \times p\) orthonormal. Then we have \(\|X - XBA^T\|_F^2 = \sum_{j=1}^k \|x_j - X\beta_j\|_F^2 + \|XA_\perp\|_F^2\). Suppose that \(A\) is given, then the optimal \(B\) can be obtained by minimizing \(\sum_{j=1}^k \|x_j - X\beta_j\|_F^2 + \lambda \sum_{j=1}^k \|\beta_j\|_2^2\), which is equivalent to \(k\) independent ridge regression problems.

3 | THE METHODS

3.1 | Feature grouping and sparse loadings

In order to investigate the structures among loadings, we extend the optimization problem (2) by imposing feature grouping and feature selection penalties simultaneously, in order to get feature grouping and sparse loadings simultaneously. The proposed FGSPCA model is based on solving the following optimization problem,

\[
\min_{A, B} \sum_{i=1}^n \|x_i - AB^T x_i\|_F^2 + \lambda_1 \sum_{j=1}^k \|\beta_j\|_2^2 + \lambda_2 \sum_{j=1}^k p_1(\beta_j) + \lambda_3 \sum_{j=1}^k p_2(\beta_j), \text{s.t. } A^TA = I_{k \times k},
\]

where \(p_1(\beta)\) and \(p_2(\beta)\) are regularization functions, taking the following penalty forms,

\[
p_1(\beta_j) = \max_{i \in E} \left\{ \frac{|\beta_j|}{\tau} - 1 \right\}, \quad p_2(\beta_j) = \sum_{i \in E} \min \left\{ \frac{|\beta_j| - \beta_j^*}{\tau}, 1 \right\},
\]

where \(\beta_j^* = \frac{\sum_{i \in E} \min \left\{ \frac{|\beta_j| - \beta_j^*}{\tau}, 1 \right\}}{\sum_{i \in E} \min \left\{ \frac{|\beta_j| - \beta_j^*}{\tau}, 1 \right\}}\).
where $\beta_{(j)}$ denotes the $j$-th element of the vector $\beta$, $\lambda_1 > 0$ and $\lambda_2 > 0$ are the corresponding tuning parameters, and $r > 0$ is a thresholding parameter, which determines when a small coefficient or a small difference between two coefficients will be penalized. The notation $E \subseteq \{1, \ldots, p\}^2$ refers to a set of edges on a fully connected and undirected graph (complete graph), with $l \neq l'$ indicating an edge directly connecting two distinct nodes $l \neq l'$, where each node represents a variable. Figure 1 gives a comparison of different penalty functions and their thresholding functions. We refer the reader to Appendix B in Supporting Information for more structured sparsity regularization functions.

Remark 1. (A) The key point of the FGPCA with $p_1(\cdot)$ and $p_2(\cdot)$ penalty functions can be viewed as performing feature selection and feature grouping simultaneously. (B) As shown in Shen et al. (2012), the truncated $\ell_1$-function $\min(\frac{\lambda}{\tau}, 1)$ can be regarded as a non-convex and non-smooth surrogate of $\ell_0$-function $I(\beta_{(j)} \neq 0)$ when $\tau \rightarrow 0$. Besides, the selection consistency can be achieved by the $\ell_0$-penalty and its surrogate—the truncated $\ell_1$-penalty (Dai et al., 2021; Shen et al., 2013). Therefore, the sparse PCA with the $\ell_1$ penalty cannot achieve selection consistency. The intuition is that, compared with the $\ell_1$ penalty, the truncated $\ell_1$ penalty is closer to the $\ell_0$ penalty and penalizes more aggressively with small coefficients preferred. Meanwhile, the truncated $\ell_1$-function $\min(\frac{\lambda}{\tau}, 1)$ can be a good approximation of the $\ell_1$-function as $\tau \rightarrow \infty$. (C) One may use the $\ell_1$-function $|\beta_{(j)}|$ as a smooth approximation of the $\ell_0$-function. However, the shrinkage bias tends to be larger as the parameter size gets larger (Wu et al., 2018; Yun et al., 2019) since the $\ell_1$ penalty is proportional to the size of parameters. The smooth approximation, $\ell_1$-function, has the drawback of producing biased estimates for large coefficients and lacking the oracle property (Fan & Li, 2001; Zhang & Huang, 2008).

3.2 Connection to sparse PCA variants

By relaxing the orthogonality requirement and extending the elastic net procedure, the sparse PCA (SPCA) (Zou et al., 2006) solves the following regularized optimization problem,

$$
(A, B) = \arg \min_{A, B} \sum_{i=1}^{n} \left( |x_i - AB^T x_i|_2^2 + \lambda_1 \sum_{j=1}^{k} |\beta_{(j)}|_2^2 + \lambda_2 \sum_{j=1}^{k} |\beta_{(j)}|_1 \right) \text{subject to } A^T A = I_{k \times k}.
$$

Note that the optimization problem in (5) is a special case of (3) as $r \rightarrow \infty$ and $\lambda_2 = 0$. By imposing a fusion penalty, the sparse fused PCA (SFPCA) with sparse fused loadings (Guo et al., 2010) solves the following regularized optimization problem,

$$
\min_{A, B} \sum_{i=1}^{n} \left( |x_i - AB^T x_i|_2^2 + \lambda_1 \sum_{j=1}^{k} |\beta_{(j)}|_2^2 + \lambda_2 \sum_{j=1}^{k} |\beta_{(j)}|_1 \right) + \lambda_2 \sum_{j=1}^{k} \sum_{l \neq t} |p_{(j,t)}| \text{subject to } A^T A = I_{k \times k},
$$

where $p_{(j,t)}$ denotes the sample correlation between variables $X_j$ and $X_t$, and $\text{sign}(x)$ returns the sign of $x$. For fair comparison, we add an $\ell_2$ penalty to the objective function of the SFPCA criterion. The SFPCA (Guo et al., 2010) can obtain sparse fused loadings in a more interpretable way, where

**Figure 1** Comparison of different penalty functions (left panel): the $\ell_1$-function (solid line), the truncated $\ell_1$-function (dashed line), and the $\ell_0$-function (dotted line), and their corresponding thresholding functions (right panel). The truncated $\ell_1$-function $\min(\frac{\lambda}{\tau}, 1)$ approximates the $\ell_0$-function $I(x \neq 0)$ as $\tau \rightarrow 0$, and it is closer to the $\ell_0$ penalty than the $\ell_1$ penalty. The thresholding functions show that, compared with the $\ell_1$ penalty, the truncated $\ell_1$ penalty penalizes more aggressively with small coefficients preferred, and it has no bias with large coefficients.
the fusion penalty depends on the sample correlation, serving as prior knowledge. Therefore, the SFPCA encourages the loadings associated with high correlation to have the same magnitude.

4 | THE ALGORITHMS

4.1 Alternating optimization algorithm of FGSPCA

In this section, we discuss the algorithms to optimize the proposed objective function in (3). An alternating optimization algorithm over A and B is employed, analogously to the SPCA algorithm (Zou et al., 2006) and SFPCA algorithm (Guo et al., 2010). Specially, the alternating algorithm to solve the optimization problem (3) proceeds as follows.

Algorithm 1. The FGSPCA algorithm.

Step 1. Initialize A by setting it to be V, the first k ordinary PC vectors.

Step 2. (Estimation of B given A). Given a fixed \(A = [\alpha_1, ..., \alpha_k]\), minimizing the objective function (3) over B is equivalent to solving the following k separate subproblems, for \(j = 1, ..., k,\)

\[
\hat{\beta}_j = \arg\min_{\beta_j} \|Z_j - X\beta_j\|^2 + \lambda_1 \sum_{i=2}^p \min_{|t|} \left( \frac{|y_{ij}^t|}{t}, 1 \right) + \lambda_2 \sum_{i \in \mathcal{F} \setminus \{j\} \in \mathcal{E}} \min_{|t|} \left( \frac{|y_{ij}^t - \hat{y}_{ij}^t|}{t}, 1 \right),
\]

where \(Z_j = X\alpha_j\). The optimization of (7) is discussed in Section 4.2. In this step, we update B and obtain the estimate \(\hat{B} = [\hat{\beta}_1, ..., \hat{\beta}_k]\).

Step 3. (Estimation of A given B). Given a fixed \(B = [\beta_1, ..., \beta_k]\), minimizing the objective function (3) over A is equivalent to solving the following problem,

\[
\min_A \|X - X\beta A\|^2 \quad \text{s.t.} \quad A^T A = I_{k \times k}.
\]

The solution to (8) can be obtained through a reduced rank Procrustes Rotation (Theorem 4 in Zou et al., 2006), see Appendix C in Supporting Information). We compute the SVD of \(X^T X^B\) as \(X^T X^B = UDV^T\), and then the solution of (8) is derived by \(A = UV^T\). In this step, we update A and obtain the estimate \(\hat{A} = [\hat{\alpha}_1, ..., \hat{\alpha}_k]\).

Step 4. Repeat Steps 2–3 until convergence.

Remark 2. (A) The initialization of \(A, V\), can be loadings of any PCA method. For simplicity, let \(V\) be the first \(k\) ordinary PC loadings. Clearly, \(V\) can also be initialized as the first \(k\) PC loadings of SPCA (Zou et al., 2006), or the first \(k\) PC loadings of SFPCA (Guo et al., 2010). (B) The convergence criterion in Step 4 can be verified by that the difference between two adjacent iterations of \(B\) is small.

We use the Frobenius norm to measure the matrix difference, that is, \(\|B_1 - B_2\|^2 \leq \epsilon\), where \(\epsilon\) is a small positive value, say, 1e−5.

4.2 Estimation of B given A

Efficiently solving the subproblem (7) plays a key role in solving the problem (3). The objective function (7) is a special case of a regularized regression problem with feature grouping and sparsity constraints (FGS). Thus, this section gives an algorithm for the FGS problem, which is a core part of Algorithm 1. The general form of the FGS problem is stated as follows:

\[
\min_{\beta_j} \sum_{i=1}^n (y_i - X^t \beta_j)^2 + \lambda_1 \sum_{i=1}^p \beta_i + \lambda_2 \sum_{i \in \mathcal{F} \setminus \{j\} \in \mathcal{E}} \min_{|t|} \left( \frac{|y_{ij}^t|}{t}, 1 \right),
\]

Since the above problem (9) is a non-convex optimization problem, we employ the difference-of-convex programming (DC) (An & Tao, 2005). Our algorithmic solution for (9) is an extension of the algorithms in Qin et al. (2020) and Shen et al. (2012) by adding the \(\epsilon_2\) penalty. Our main technical contribution is to extend the algorithm in Qin et al. (2020) and Shen et al. (2012) to applications of developing more interpretable PCA.
We propose an integrated algorithm for the estimation of \( B \) given \( A \) (Algorithm 1), which integrates the difference-of-convex algorithm (DC), the augmented Lagrange method and the coordinate descent method (AL-CD), for efficient computation. The procedure to solve the FGS problem consists of three steps. First, the non-convex objective function is decomposed into a difference of two convex functions using DC. Then a sequence of approximations of the trailing convex function is constructed with its affine minorization (through linearizing). Second, a quadratic constrained optimization problem is solved via the coordinate descent method. The detailed derivation procedures of DC, AL-CD are given in Appendix D in Supporting Information. For simplicity, only the derived results are provided.

Denote \( \beta_g = \beta - \beta_f \) and define \( \xi = (\beta_1, \ldots, \beta_p, \beta_{12}, \ldots, \beta_{1p}, \ldots, \beta_{p-1}p \} \). Then update \( \hat{\beta}^{(m)} \) by the following formulas, for \( k = 1, 2, \ldots \):

- Given \( \hat{\beta}^{(m-k-1)}_l \), update \( \hat{\beta}^{(m-k)}_l \) by
  \[
  \hat{\beta}^{(m-k)}_l = \alpha^{-1} \gamma,
  \]
  where \( \alpha = 2d + 2 \sum_{j=1}^p x_j^2 + \nu(k) \cdot \|l\| \in \mathcal{E}(m-1) \). And \( \gamma = \gamma^* \) if \( |\hat{\beta}^{(m-k)}_l| \geq \tau \); otherwise, \( \gamma = ST(\gamma^* \frac{1}{\rho}) \). Here, \( ST(x, \delta) = \text{sign}(x)(|x| - \delta)_+ \) is the soft threshold function, and

- Given \( \hat{\beta}^{(m-k-1)}_l \), update \( \hat{\beta}^{(m-k)}_l \) (1 \( \leq l \leq p \)) (with \( \hat{\beta}^{(m-k)}_l \) already updated and fixed). Then

\[
\hat{\beta}^{(m-k)}_l = \begin{cases} 
\frac{1}{\rho(k)} \text{ST} \left( \frac{\beta_l^{(k)} + \nu(k)}{\beta_l^{(m-k)} - \hat{\beta}^{(m-k)}_l}, \frac{\lambda_2}{\tau} \right) \text{ if } (l, k) \in \mathcal{E}(m-1), \\
\hat{\beta}^{(m-k)}_l \text{ if } (l, k) \notin \mathcal{E}(m-1).
\end{cases}
\]

The process of coordinate descent iterates until convergence, satisfying the termination condition \( \|\hat{\beta}^{(m)} - \hat{\beta}^{(m-k-1)}\|_\infty \leq \delta^* \) (e.g., \( \delta^* = 10^{-5} \)). Hence, \( \hat{\beta}^{(m)} = \hat{\beta}^{(m-k)} \), where \( t^* \) denotes the iteration at termination. Specially, we take \( \rho = 1.05, \nu = 1, \delta^* = 10^{-5} \) in the simulations.

### 4.3 Convergence and computational complexity

The convergence of the algorithm essentially follows the standard result. Note that we have a closed-form solution of \( A \) when fixing \( B \). Since the truncated penalties are not convex in \( B \), and thus the objective function is not convex in \( B \) when fixing \( A \), and that is when the difference-of-convex function kicks in to convert the non-convex function to the difference of two convex functions. When solving the problem (7), the proposed algorithm could potentially lead to a local optimum as the objective function of estimating \( B \) when fixing \( A \) in (7) is non-convex. But the objective function with linear constraints in the AL-CD procedure obtained from the local linear approximation is differentiable everywhere, and thus, the convergence of coordinate descent is guaranteed. Therefore, it is only necessary to ensure that each step is guaranteed to converge. In Step 3, the optimized objective function is (8), and we can obtain the exact solution in the closed form. In Step 2, we solve the optimization problem (7) iteratively. The convergence of the integrated algorithm for the subproblem of estimating \( B \) when fixing \( A \) is given in Lemma 2. Denote

\[
S(\beta) = \sum_{i=1}^n (y_i - x_i \beta)^2 + \sum_{i=1}^p \|\beta_i\|^2 + \lambda_1 \sum_{i=1}^p \min \left( \frac{|\beta_i|}{\tau}, 1 \right) + \lambda_2 \sum_{l \in \mathcal{E}} \min \left( \frac{|\beta_l - \hat{\beta}^{(m-k)}_l|}{\tau}, 1 \right).
\]

**Lemma 2.** The proposed algorithm for estimation of \( B \) given \( A \) converges. That is,

\[
S(\hat{\beta}^{(m)}) \to c, \text{ as } m \to \infty,
\]

where \( c \) is a constant value and \( m \) is the number of iterations of the integrated algorithm for problem (9).
Lemma 2 above guarantees the convergence of the algorithm for estimation of $B$ given $A$ theoretically, which is analogous to Theorem 1 in Shen et al. (2012) and Theorem 3 in Qin et al. (2020). Thus, we omit the proof. It is crucial to pick a suitable initial value $\hat{\beta}^{(0)}$. Since (7) is a regression problem, possible candidate initial values are those estimated by any regression solver, such as glmnet in R and sklearn in python.

As for the computational complexity, the coordinate descent updating involves calculation of $\sum_{i=1}^{n}x_{il}^2$ and $\sum_{i=1}^{n}x_{il}y_{il}$, which requires $O(np^2k)$ operations. The construction of $\tilde{z}$ requires $O(kp^2)$ operations. Therefore, each update in updating $B$ is of order $O(np^2k)$. The estimation of $A$ by solving an SVD needs $O(pk^2)$ operations. The total computational cost is $O(np^2k) + O(pk^2)$.

4.4 | The selection of tuning parameters

Cross-validation (CV) is one way to select the optimal values, but it is computationally expensive. Here, the Bayesian information criterion (BIC) is employed as the approach for tuning parameter selection, which we use in simulations in Section 5. In general, solutions from cross-validation and BIC are comparable. We select the model that has the minimum BIC value when using such criteria. Our proposed method has four tuning parameters, $\lambda_1, \lambda_2, \tau, r$. Let $\phi$ denote the parameters that need to be tuned or selected in the candidate model. Then $\psi = (\lambda_1, \lambda_2, \tau, r)$ for our proposed FGSPCA, $\phi = (\lambda_1, \lambda_2)$ for SFPCA (Guo et al., 2010), and $\phi = (\lambda, \tau)$ for the SPCA (Zou et al., 2006). Let $A^\phi = [a_1^\phi, ..., a_k^\phi]$, and $B^\phi = [b_1^\phi, ..., b_k^\phi]$, be the estimates of $A$ and $B$ in (3) based on the tuning parameters $\phi$.

We define the BIC criterion of PCA variants as follows:

$$\text{BIC}(\psi) = n \log \left( \frac{\|X - XB^\phi(A^\phi)^T\|_F^2}{n} \right) + \log(n) \cdot df,$$

where $df$ represents the degree of freedom, denoted $df^{\text{SPCA}}$ for SPCA (Zou et al., 2006), $df^{\text{SFPCA}}$ for SFPCA (Guo et al., 2010), and $df^{\text{FGSPCA}}$ for FGSPCA. Specially, $df^{\text{SPCA}}$ is defined as the number of all nonzero elements in $B^\psi$, and $df^{\text{SFPCA}}$ and $df^{\text{FGSPCA}}$ are defined as the number of all non-different groups in $B^\psi$. The definitions are similar to $df$ defined for Lasso and fused Lasso (Tibshirani et al., 2005; Zou et al., 2007). Intuitively, the involvement of the truncated parameter $r$ makes more complex the method and the parameter tuning process. However, empirical studies show that the involvement of the truncated parameter $r$ establishes a trade-off between $r$ and $(\lambda_1, \lambda_2)$, reducing the sensitivity of the tuning of $\lambda_1$ and $\lambda_2$.

5 | EXPERIMENTS

5.1 | Adjusted variance

Denote $\tilde{Z} = [\tilde{Z}_1, ..., \tilde{Z}_k]$ the modified PCs. Due to the grouping and sparsity constraints, $\tilde{Z}_i$ is no longer orthogonal to $\tilde{Z}_i, i = 1, ..., k - 1$. Instead, they are correlated with each other. Thus, we remove from $\tilde{Z}_i$ the correlation effect of $\tilde{Z}_i, i = 1, ..., k - 1$ using regression projection. The definition of the adjusted variance is adopted from Zou et al. (2006), which is computed based on the QR decomposition. Suppose $\tilde{Z} = QR$, where $Q$ is orthogonal and $R$ is upper triangular. The adjusted variance of the $j$-th PC is $R_{jj}^2$. The explained total variance is the cumulative adjusted variance, which is defined as $\sum_{j=1}^{k} R_{jj}^2$.

5.2 | Pitprops data

The pitprops data are a classic dataset widely used for PCA analysis, as it is usually difficult to show the interpretability of principal components. In the pitprops data, there are 180 observations and 13 measured variables. It is used in ScoTLASS (Jolliffe et al., 2003) and SPCA (Zou et al., 2006). As a demonstration of the performance of the FGSPCA method, especially the grouping effect and sparsity effect, we consider the first six PCs of pitprops data.

Table 1 shows the sparse loadings and the corresponding variance obtained by SPCA (Zou et al., 2006) and FGSPCA. As can be seen from Table 1, both SPCA and FGSPCA show strong sparsity effects with respect to the number of zero loadings. On the other hand, FGSPCA has a stronger grouping effect in terms of the number of loading groups, while SPCA has a weaker grouping effect compared with FGSPCA. Interestingly, through the grouping effect introduced in FGSPCA, FGSPCA shows a stronger sparsity compared with SPCA with respect to the number of zeroes. In detail, for the first PC obtained by FGSPCA, the loadings belong to two distinct groups with nonzero values and one sparse group with zero values. Furthermore, these groups are learned automatically from the FGSPCA model rather than from prior knowledge. The grouping effects among loadings further improve the interpretability of the PCA.
It can be seen from Figure 2 that the first six PCs obtained by FGSPCA and SPCA account for almost the same amount of total variance, 74.96% for FGSPCA and 75.77% for SPCA, respectively, which is much larger compared with SCoTLASS (69.3%). The significant improvement in the total variance explained by FGSPCA and SPCA may result from the sparse structure on the loadings, since the derived PCs obtained by SCoTLASS are not sparse enough as analyses in Zou et al. (2006).
5.3 | Synthetic data

5.3.1 | Simulation 1

We adopt the same synthetic example settings as Zou et al. (2006). The generating mechanism of the synthetic data consists of three hidden factors, that is,

\[
\begin{align*}
V_1 &\sim N(0,290), \\
V_2 &\sim N(0,300), \\
V_3 &= -0.3V_1 + 0.925V_2 + \epsilon, \ 
&\epsilon \sim N(0,1),
\end{align*}
\]

(13)

where \(V_1, V_2, \epsilon\) are independent. Next, 10 observable variables are constructed as follows:

\[
X_j = \begin{cases} 
V_1 + \epsilon, & 1 \leq j \leq 4, \\
V_2 + \epsilon, & 5 \leq j \leq 8, \\
V_3 + \epsilon, & j = 9,10,
\end{cases}
\]

where \(\epsilon_j, j = 1,\ldots,10\), are independent and identically distributed (i.i.d) with \(N(0,1)\). Note that the variances of the three hidden factors are 290, 300, and 282.7875, respectively. Note that by the data generating mechanism, the variables \(X_1\) to \(X_4\) form a block/group with a constant weight (“block 1”), while variables \(X_5\) to \(X_8\) and \(X_9, X_{10}\) form another two blocks, “block 2” and “block 3,” respectively. Since “block 2” and “block 3” are highly correlated, thus, they can be merged into one group, say “BLOCK 0.” Ideally, a sparse first derived PC1 should recover “BLOCK 0” of the hidden factor \(V_2\) using \(X_5, X_6, X_7, X_8, X_9, X_{10}\) with equal loadings, while a sparse second derived PC2 should pick up \(X_1, X_2, X_3, X_4\) to recover “block 1” of the hidden factor \(V_1\) with the same weights, since variance of \(V_2\) is larger than that of \(V_1\).

Zou et al. (2006) computed sparse PCA using the true covariance matrix as the data generating mechanism is known and the true covariance matrix of the 10 observable variables \(\{X_1,\ldots,X_{10}\}\) can be easily calculated. In our simulation, we adopt the same setting and the procedure as in Guo et al. (2010) to generate data \(X_{n,p}\) with \(n = 50\) according to the above data generating mechanism and repeated the simulation 50 times. And we perform the ordinary PCA, SPCA (sparse PCA), ST (simple thresholding), and FGSPCA on \(X\). The PC loadings from ordinary PCA, SPCA, ST, and FGSPCA are reported in Table 2.

| Variable | PCA PC1 | PCA PC2 | PCA PC3 | SPCA PC1 | SPCA PC2 | ST PC1 | ST PC2 | FGSPCA PC1 | FGSPCA PC2 |
|----------|---------|---------|---------|-----------|-----------|--------|--------|------------|------------|
| X1       | 0.116   | 0.479   | 0.062   | 0.5       | -0.5      | 0.5    |        |            |            |
| X2       | 0.116   | 0.479   | 0.059   | 0.5       | -0.5      | 0.5    |        |            |            |
| X3       | 0.116   | 0.479   | 0.114   | 0.5       | -0.5      | 0.5    |        |            |            |
| X4       | 0.116   | 0.479   | 0.114   | 0.5       | -0.5      | 0.5    |        |            |            |
| X5       | -0.395  | 0.145   | -0.269  | -0.5      | -0.408    |        |        |            |            |
| X6       | -0.395  | 0.145   | -0.269  | -0.5      | -0.408    |        |        |            |            |
| X7       | -0.395  | 0.145   | -0.269  | -0.5      | -0.497    | -0.408 |        |            |            |
| X8       | -0.395  | 0.145   | -0.269  | -0.5      | -0.497    | -0.408 |        |            |            |
| X9       | -0.401  | -0.010  | 0.582   | -0.503    | -0.408    |        |        |            |            |
| X10      | -0.401  | -0.010  | 0.582   | -0.503    | -0.408    |        |        |            |            |
| No. Groups | 3 3 5 | 1 1 2 | 1 1 |            |            |        |        |            |            |
| No. Nonzeros | 10 10 10 | 4 4 4 | 6 4 |            |            |        |        |            |            |
| Variance (%) | 69.64 30.36 | 41.02 39.65 | 38.88 39.65 | 59.01 39.65 |        |        |            |            |
| Adj. V (%) | - - - | 41.02 39.65 | 38.88 38.73 | 59.08 39.25 |        |        |            |            |
| CV (%) | 69.64 100 100 | 41.02 80.67 | 38.88 77.61 | 59.08 98.33 |        |        |            |            |

Note: “Adj. V (%)” is the proportions of adjusted variance, and “CV (%)” is the proportions of cumulative adjusted variance.
Table 2 lists three PCs of the ordinary PCA. It shows that the first two PCs account for 100% of the total explained variance, suggesting that other dimension reduction methods can consider only the first two derived PCs. Results on Table 2 show that all the methods (SPCA, ST, and FGSPCA) can perfectly recover “block 1” with the hidden factor $V_1$ using the derived PC2. However, as for the first derived PC1, SPCA recovers the hidden factor $V_2$ only using $X_5, X_6, X_7, X_8$ without $X_9, X_{10}$, as the weights on $X_9, X_{10}$ are zeroes. The ST method recovers the hidden factor $V_2$ using $X_7, X_8, X_9, X_{10}$, which is far from being correct by imposing zero weights on $X_9, X_{10}$. FGSPCA perfectly recovers the hidden factor $V_2$ using $X_5, X_6, X_7, X_8, X_9, X_{10}$ with the same weights, which is consistent with the ideal results analysed above.

The results of variance from Table 2 show that the total variance explained by the first two PCs is 98.33% for FGSPCA and 80.67% for SPCA, a great improvement of 17.66% due to the grouping effects of FGSPCA. Moreover, compared with the ordinary PCA (100% explained total variance), FGSPCA is only 1.67% less with respect to the total variance explained. Most importantly, FGSPCA achieves a remarkable improvement in the interpretability of PCs with the same value, which is the grouping effect.

5.3.2 | Simulation 2

In this example, we consider a high dimensional version ($p > n$) of Simulation 1. We define

$$X_j = \begin{cases} 
V_1 + \epsilon_j, & 1 \leq j \leq 20, \\
V_2 + \epsilon_j, & 21 \leq j \leq 40, \\
V_3 + \epsilon_j, & 41 \leq j \leq 50,
\end{cases}$$

where $\epsilon_j, j = 1, \ldots, 50$, are i.i.d. $N(0,1)$. We generate a data matrix $X_{n \times p}$ with $n = 20$, and we conduct 50 repetitions. The estimated loadings are illustrated in Figure 3. Results show that SFPCA and FGSPCA produce similar sparse structures in the loadings. However, compared with the ‘scattered’ loadings from SFPCA, the loadings estimated by FGSPCA are smooth and easier for interpretation.

![Figure 3](image-url)  
**Figure 3** Factor loadings of the first (left column) and second (right column) PC vectors estimated by SFPCA (Guo et al., 2010) (first row) and our proposed FGSPCA (second row). The horizontal axis is the variables and the vertical axis is the value of the loadings. Each colored curve represents the PC vector in one replication.
6 | DISCUSSION AND EXTENSION

One limitation of FGSPCA is that it uses non-convex regularizers, neither smooth nor differentiable. Recent research work (Birgin et al., 2021; Wen et al., 2018; Zhang et al., 2020) has shown better denoising advantages of non-convex regularizers over convex ones. However, when solving the subproblem (7) with non-convex penalties, the proposed method could potentially lead to a local optimum, as the objective function in (7) is non-convex. As is pointed out in Wen et al. (2018), the performance of non-convex optimization problems is usually closely related to the initialization, which are inherent drawbacks of non-convex optimization problems. Hence, it is desirable to pick a suitable initial value of \( \hat{\beta}^{(0)} \). Since each subproblem (7) is a classical regression problem, possible candidate initial values are those estimated by any regression solver, such as the R package glmnet (Friedman et al., 2009) and the python scikit-learn. For simplicity, we use the result of SVD as the initialization in this paper.

The FGSPCA can be easily extended to the case with non-negative loadings, namely nnFGSPCA. In light of the work in Qin et al. (2020), we incorporate another regularization term, \( p_3(\beta) = \sum_{j=1}^{p} (\min(\beta_j, 0))^2 \) that characterizes the non-negativity, into the objective function. The optimization problem of nnFGSPCA becomes

\[
\min_{A,B} \sum_{i=1}^{n} \| x_i - AB^T x_i \|^2 + \lambda_1 \sum_{j=1}^{k} \| \beta_j \|^2 + \lambda_2 \sum_{j=1}^{k} p_1(\beta_j) + \lambda_3 \sum_{j=1}^{k} p_2(\beta_j) + \lambda_4 \sum_{j=1}^{k} p_3(\beta_j),
\]

s.t. \( A^T A = I_{p \times k} \).

The nnFGSPCA can be easily solved using similar techniques (see Appendix E in Supporting Information for details).

7 | CONCLUSION

In this paper, we propose the FGSPCA method to produce modified principal components by considering additional grouping structures where the loadings share similar coefficients (i.e., feature grouping), besides a special group with all coefficients being zero (i.e., feature selection). The proposed FGSPCA method can perform simultaneous feature clustering/grouping and feature selection by imposing the non-convex regularization with naturally adjustable sparsity and grouping effects. Therefore, the model learns the grouping structure rather than from given prior information. Efficient algorithms are designed and experiment results show that the proposed FGSPCA benefits from the grouping effect compared with methods without grouping effect.

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in R package elasticnet at https://rdrr.io/cran/elasticnet/man/pitprops.html. These data were derived from the following resources available in the public domain: - R, https://rdrr.io/cran/elasticnet/man/pitprops.html.

ORCID

Haiyan Jiang https://orcid.org/0000-0002-4099-480X

REFERENCES

An, L. T. H., & Tao, P. D. (2005). The DC (difference of convex functions) programming and DCA revisited with dc models of real world nonconvex optimization problems. *Annals of Operations Research*, 133(1), 23–46.

Birgin, E. G., Martínez, J. M., & Ramos, A. (2021). On constrained optimization with nonconvex regularization. *Numerical Algorithms*, 86(3), 1165–1188.

Cai, T. T., Ma, Z., & Wu, Y. (2013). Sparse PCA: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6), 3074–3110.

Croux, C., Filzmoser, P., & Fritz, H. (2013). Robust sparse principal component analysis. *Technometrics*, 55(2), 202–214.

Dai, M., Dai, S., Huang, J., Kang, L., & Lu, X. (2021). Truncated L1 regularized linear regression: Theory and algorithm. *Communications in Computational Physics*, 30(1), 190–209.

Erichson, N. B., Zheng, P., Manohar, K., Brunton, S. L., Kutz, J. N., & Aravkin, A. Y. (2020). Sparse principal component analysis via variable projection. *SIAM Journal on Applied Mathematics*, 80(2), 977–1002.

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

Friedman, J., Hastie, T., & Tibshirani, R. (2009). glmnet: Lasso and elastic-net regularized generalized linear models. *R Package Version*, 1(4), 1–24.

Grbović, M., Dance, C., & Vucetic, S. (2012). Sparse principal component analysis with constraints. In *Proceedings of the aais conference on artificial intelligence*, 26.

Guo, J., James, G., Levina, E., Michailidis, G., & Zhu, J. (2010). Principal component analysis with sparse fused loadings. *Journal of Computational and Graphical Statistics*, 19(4), 930–946.
Jenatton, R., Audibert, J.-Y., & Bach, F. (2011). Structured variable selection with sparsity-inducing norms. The Journal of Machine Learning Research, 12, 2777–2824.

Jenatton, R., Obozinski, G., & Bach, F. (2010). Structured sparse principal component analysis. In Proceedings of the thirteenth international conference on artificial intelligence and statistics (pp. 366–373). JMLR Workshop and Conference Proceedings.

Jin, Y., & Sidford, A. (2019). Principal component projection and regression in nearly linear time through asymmetric SVRG. In Advances in neural information processing systems, 32, Curran Associates, Inc.

Johnstone, I. M., & Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. Journal of the American Statistical Association, 104(486), 682–693.

Jolliffe, I. T. (1986). Principal components in regression analysis, Principal component analysis (pp. 129–155). Springer.

Jolliffe, I. T., Trendafilov, N. T., & Uddin, M. (2003). A modified principal component technique based on the lasso. Journal of Computational and Graphical Statistics, 12(3), 531–547.

Khan, Z., Shafait, F., & Mian, A. (2015). Joint group sparse PCA for compressed hyperspectral imaging. IEEE Transactions on Image Processing, 24(12), 4934–4942.

Nadler, B. (2008). Finite sample approximation results for principal component analysis: A matrix perturbation approach. The Annals of Statistics, 36(6), 2791–2817.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. Statistica Sinica, 2007, 1617–1642.

Qin, S., Ding, H., Wu, Y., & Liu, F. (2020). High-dimensional sign-constrained feature selection and grouping. Annals of the Institute of Statistical Mathematics, 2020, 1–33.

Shen, X., Huang, H.-C., & Pan, W. (2012). Simultaneous supervised clustering and feature selection over a graph. Biometrika, 99(4), 899–914.

Shen, X., Pan, W., Zhu, Y., & Zhou, H. (2013). On constrained and regularized high-dimensional regression. Annals of the Institute of Statistical Mathematics, 65(5), 807–832.

Tian, L., Nie, F., Wang, R., & Li, X. (2020). Learning feature sparse principal subspace. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., & Lin, H. (Eds.), Advances in neural information processing systems (Vol. 33, pp. 14997–15008). Curran Associates, Inc.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(1), 91–108.

Vu, V. Q., Cho, J., Lei, J., & Rohe, K. (2013). Fantope projection and selection: A near-optimal convex relaxation of sparse PCA. In Proceedings of the 26th international conference on neural information processing systems - volume 2 (pp. 2670–2678). NIPS’13 Curran Associates Inc., Red Hook, NY, USA.

Wang, Z., Lu, H., & Liu, H. (2014). Tighten after relax: Minimax-optimal sparse PCA in polynomial time. Advances in Neural Information Processing Systems, 2014, 3383.

Wen, F., Chu, L., Liu, P., & Qiu, R. C. (2018). A survey on nonconvex regularization-based sparse and low-rank recovery in signal processing, statistics, and machine learning. IEEE Access, 6, 69883–69906.

Wu, C., Liu, Z., & Wen, S. (2018). A general truncated regularization framework for contrast-preserving variational signal and image restoration: Motivation and implementation. Science China Mathematics, 61(9), 1711–1732.

Yi, S., Lai, Z., He, Z., Cheung, Y., & Liu, Y. (2017). Joint sparse principal component analysis. Pattern Recognition, 61, 524–536.

Yun, J., Zheng, P., Yang, E., Lozano, A., & Aravkin, A. (2019). Trimming the $\ell_1$ regularizer: Statistical analysis, optimization, and applications to deep learning. In International conference on machine learning (pp. 7242–7251). PMLR.

Zhang, C.-H., & Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. The Annals of Statistics, 36(4), 1567–1594.

Zhang, R., & Tong, H. (2019). Robust principal component analysis with adaptive neighbors. In Advances in neural information processing systems, 32, Curran Associates, Inc.

Zhang, Y., Li, S., Guo, Z., & Wu, B. (2020). An adaptive total variational despeckling model based on gray level indicator frame. Inverse Problems & Imaging, 2020, 1.

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.

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. Journal of Computational and Graphical Statistics, 15(2), 265–286.

Zou, H., Hastie, T., & Tibshirani, R. (2007). On the degrees of freedom of the lasso. The Annals of Statistics, 35(5), 2173–2192.

Zou, H., & Xue, L. (2018). A selective overview of sparse principal component analysis. Proceedings of the IEEE, 106(8), 1311–1320.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Jiang, H., Qin, S., & Madrid Padilla, O. H. (2023). Feature grouping and sparse principal component analysis with truncated regularization. Stat, 12(1), e538. https://doi.org/10.1002/sta4.538