Journal of Multivariate Analysis 166 (2018) 1–16

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Integrative sparse principal component analysis

Kuangnan Fang [a,b], Xinyan Fan [a], Qingzhao Zhang [a,b,c], Shuangge Ma [d,*]

[a] Department of Statistics, School of Economics, Xiamen University, China
[b] Fujian Key Laboratory of Statistical Science, Xiamen University, China
[c] The Wang Yanan Institute for Studies in Economics, Xiamen University, China
[d] Department of Biostatistics, Yale School of Public Health, USA

ARTICLE INFO

Article history:
Received 2 September 2017
Available online 12 February 2018

AMS subject classifications:
62H25

Keywords:
Contrasted penalization
Integrative analysis
Sparse PCA

ABSTRACT

In the analysis of data with high-dimensional covariates and small sample sizes, dimension reduction techniques have been extensively employed. Principal component analysis (PCA) is perhaps the most popular dimension reduction technique. To remove noise effectively and generate more interpretable results, the sparse PCA (SPCA) technique has been developed. In high dimension, the analysis of a single dataset often generates unsatisfactory results. In a series of studies under the "regression analysis + variable selection" setting, it has been shown that integrative analysis provides an effective way of pooling information from multiple independent datasets and outperforms single-dataset analysis and many alternative multi-datasets analyses, especially including the classic meta-analysis. In this study, with multiple independent datasets, we propose conducting dimension reduction using a novel iSPCA (integrative SPCA) approach. Penalization is adopted for regularized estimation and selection of important loadings. Advancing from the existing integrative analysis studies, we further impose contrasted penalties, which may generate more accurate estimation/selection. Multiple settings on the similarity across datasets are comprehensively considered. Consistency properties of the proposed approach are established, and effective computational algorithms are developed. A wide spectrum of simulations demonstrate competitive performance of iSPCA over the alternatives. Two sets of data analysis further establish its practical applicability.

1. Introduction

Data with high-dimensional covariates and small-to-moderate sample sizes abound. Extensive methodological and theoretical studies have been conducted. In particular, dimension reduction techniques such as principal component analysis (PCA), partial least squares (PLS), independent component analysis (ICA), and others, have been proposed. PCA is arguably the most popular of all. It can assist in understanding underlying data structures, clustering analysis, regression analysis, and many other tasks. We refer to [7,13,15,18] and other publications for methodological, theoretical, and numerical studies on PCA in high-dimensional settings. In many practical studies, it has been suggested that only a small subset of variables are relevant, while others are "noise". To identify relevant variables and generate more interpretable results, the sparse PCA (SPCA) technique has been developed, which applies regularized estimation to generate sparse loadings. In the literature, methodological studies on SPCA include [6,12,16,24,27], theoretical studies include [5,17], and numerical studies include [2], among others.

* Corresponding author.
E-mail address: shuangge.ma@yale.edu (S. Ma).

https://doi.org/10.1016/j.jmva.2018.02.002
0047-259X/

Despite many promising successes, it is still often observed that results generated from analyzing a single dataset are unsatisfactory. This can be partly seen from our numerical study. Although there may be multiple contributing factors, the most important is perhaps the small sample size. For many scientific problems, there are multiple independent studies with comparable settings, e.g., with the same set of variables measured on subjects with similar characteristics [3]. In a series of studies under the "regression analysis + variable selection" settings [19,26], it has been shown that integrative analysis, which jointly analyzes the raw data of multiple independent datasets, outperforms single-dataset analysis and many other multi-datasets analyses especially including the classic meta-analysis, under which each dataset is analyzed separately, and then summary statistics are pooled. Extensive methodological, theoretical, and numerical studies have been conducted [10,11].

Motivated by the importance of PCA/SPCA in the dimension reduction analysis of high-dimensional data, strong the need for improving over single-dataset analysis, and the success of integrative analysis under the "regression analysis + variable selection" setting, we present here a novel approach and conduct the integrative SPCA (iSPCA) analysis of multiple independent datasets. This study is related to, but improves upon, the existing ones in the following aspects. It is built on the SPCA technique and extends it to the integrative analysis of multiple datasets. Significant challenges arise from the "interconnections" among datasets. This study belongs to the same integrative analysis paradigm as [26], with the goal of improving estimation/selection results by pooling information from multiple independent datasets. However, in sharp contrast with the existing integrative analysis studies, dimension reduction, which is as important as variable selection, is conducted. This study can potentially pave the road to conducting integrative analysis based on other dimension reduction techniques. With these differences, significant methodological, theoretical, and numerical developments are needed beyond the existing literature. Overall, this study provides a novel and useful new venue for analyzing high-dimensional datasets.

The rest of the article is organized as follows. The proposed analysis, its theoretical properties, and computational algorithms are described in Section 2. Numerical study, including simulation and data analysis, is conducted in Section 3 to demonstrate satisfactory finite-sample performance of the proposed analysis. The article concludes with a discussion in Section 4. Additional technical details and numerical results are provided in the Appendix.

2. Methods

2.1. SPCA with a single dataset

For the analysis of a single dataset, the SPCA technique has been well developed; see, e.g., [16,24,27]. For the completeness of this article and development of notations, below we briefly describe SPCA for a single dataset and refer to the literature for more details.

Consider the data matrix $X = [X_1, \ldots, X_n]_{d\times n}$ whose n columns are iid observations from a d-dimensional multivariate normal distribution $\mathcal{N}(0, \Sigma_d)$. The covariance matrix Σ_d can be decomposed as $\Sigma_d = U_d\Lambda_d U_d^\top$, where Λ_d is the diagonal matrix of eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d$ and $U_d = [u_1, \ldots, u_d]$ is the matrix of the corresponding eigenvectors. The superscript $^\top$ denotes matrix transpose. PCA can be achieved by conducting the singular value decomposition (SVD) of X. Suppose that $\text{rank}(X) = r$, and denote the SVD of X as $X = \hat{U}_r\hat{\Omega}\hat{V}_r^\top$, where $\hat{U}_r = [\hat{u}_1, \ldots, \hat{u}_r]$, $\hat{V}_r = [\hat{v}_1, \ldots, \hat{v}_r]$, and $\hat{\Omega} = \text{diag}(\hat{\omega}_1, \ldots, \hat{\omega}_r)$. The columns of $\hat{U}_r$ are orthonormal, and so are the columns of $\hat{V}_r$. The singular values are ordered, i.e., $\hat{\omega}_1 \geq \cdots \geq \hat{\omega}_r$. With the connection between PCA and SVD, it can be easily shown that for all $i \in \{1, \ldots, r\}$, $\hat{u}_i$ is the sample estimate of eigenvector u_i, and $\hat{\omega}_i^2/n$ is the sample estimate of eigenvalue λ_i. In addition, for any integer $\ell \in \{1, \ldots, r\}$, $\tilde{X} = \sum_{i=1}^{\ell}\hat{\omega}_i\hat{u}_i\hat{v}_i^\top$ is the best rank-ℓ matrix approximation of X. That is,

$$\tilde{X} = \underset{X^*}{\text{argmin}}[\|X - X^*\|_F^2 = \text{tr}\{(X - X^*)(X - X^*)^\top\}],$$

where the subscript "F" denotes the Frobenius norm, "tr" denotes trace, and X^* has rank ℓ.

We focus on the first principal component (PC). This is equivalent to finding the best rank-1 matrix approximation of X under the Frobenius norm. Note that any $d \times n$ rank-1 matrix can be written as $\check{u}\tilde{v}^\top$, where $\tilde{v}$ is a unit-norm n-vector and $\check{u}$ is a d-vector. Thus, estimating the first PC can be formulated as the following optimization problem:

$$\underset{\check{u},\tilde{v}}{\text{argmin}}\|X - \check{u}\tilde{v}^\top\|_F^2.$$

Given the connection between PCA and SVD, we have $\check{u} = \hat{\omega}_1\hat{u}_1$ and $\tilde{v} = \hat{v}_1$.

The standard PCA generates dense loadings, making interpretation challenging under high-dimensional settings. In addition, under many practical scenarios, only a subset of variables are relevant. Motivated by such considerations, the SPCA technique has been proposed, which induces sparse loadings by applying regularized estimation. The most popular SPCA applies penalization [17,16], where the estimate is defined as

$$\underset{\check{u},\tilde{v}}{\text{argmin}}\{\|X - \check{u}\tilde{v}^\top\|_F^2 + \text{pen}(\check{u})\}, \quad \text{subject to } \|\tilde{v}\| = 1.$$

Here, pen is the penalty function, and the most popular choice is perhaps Lasso. With a sparse $\breve{u}$, the corresponding sparse loading vector is $\tilde{u} = \breve{u}/\|\breve{u}\|$.

2.2. Integrative analysis

In what follows, we conduct the integrative analysis of M independent datasets. To simplify notation, consider the scenario where all datasets have the same set of variables. As shown, e.g., in [22], the analysis can be easily adapted to accommodate the scenario where the datasets have overlapping but different sets of variables. Our goal is to conduct SPCA with all M datasets simultaneously and to enhance the identification of important loadings and estimation. Integrative analysis demands a certain level of similarity across datasets (in our case, similarity in the sets of important loadings); otherwise, information borrowing across datasets will be unlikely. The proper selection of datasets can be achieved by, e.g., analyzing meta-data; this has been discussed in [3,22] and will not be reiterated here. It is noted that, in published studies when different datasets have completely different sets of important variables [26], i.e., the worst case scenario, integrative analysis still performs competitively.

2.3. iSPCA

For each dataset $m \in \{1, \ldots, M\}$, assume that there are n_m iid observations $X_1^{(m)}, \ldots, X_{n_m}^{(m)}$ from a d-dimensional multivariate normal distribution $\mathcal{N}(0, \Sigma_d^{(m)})$. The covariance matrix $\Sigma_d^{(m)}$ can be decomposed as

$$\Sigma_d^{(m)} = U_d^{(m)} \Lambda_d^{(m)} U_d^{(m)\top},$$

where $\Lambda_d^{(m)}$ is the diagonal matrix of eigenvalues $\lambda_1^{(m)} \geq \cdots \geq \lambda_d^{(m)}$, and $U_d^{(m)} = [u_1^{(m)}, \ldots, u_d^{(m)}]$ is the matrix of the corresponding eigenvectors. Denote

$$X^{(m)} = [X_1^{(m)}, \ldots, X_{n_m}^{(m)}]_{d \times n_m}, \quad \hat{\Sigma}_d^{(m)} = X^{(m)} X^{(m)\top}/n_m.$$

In our SPCA analysis of the M datasets, as in single-dataset analysis, we focus on estimating the M first PCs. As the datasets are generated independently, differences inevitably exist. Thus, following the same spirit as in published integrative analyses, the M first PCs are not forced to be the same. Consider the objective function

$$L(\breve{u}_1^{(1)}, \ldots, \breve{u}_1^{(M)}) = \sum_{m=1}^{M} \frac{1}{2n_m} \|X^{(m)} - \breve{u}_1^{(m)} \tilde{v}_1^{(m)\top}\|_F^2 + \text{pen}(\breve{u}_1^{(1)}, \ldots, \breve{u}_1^{(M)}), \quad \text{subject to } \|\tilde{v}_1^{(1)}\| = \cdots = \|\tilde{v}_1^{(M)}\| = 1$$

where notations have similar implications as in single-dataset analysis. With independence, the first term is the sum of M individual terms. Note that the normalization by sample size is optional; it is not consistently used in published studies. Here we conduct normalization to give all datasets "equal attention".

In SPCA, the selection of relevant variables with nonzero loadings (i.e., identification of the sparsity structure of loadings) is of critical importance. With M datasets, M sparsity structures need to be considered. In integrative analysis, two generic scenarios have been considered for sparsity structures [26]. The first is the homogeneity structure, which assumes that multiple datasets have the same set of important variables (i.e., the same sparsity structure). The second is the heterogeneity structure, which allows the M sets of important variables to be possibly different. The heterogeneity structure includes the homogeneity structure as a special case and is more flexible. Thus in this study, we focus on the heterogeneity structure. For the identification of relevant variables and regularized estimation, we first consider the penalty function

$$\sum_{i=1}^{d} \rho\left(\sum_{m=1}^{M} \rho(|\breve{u}_{i,1}^{(m)}|; \mu_1, a); 1, b\right), \tag{1}$$

where $\breve{u}_{i,1}^{(m)}$ is the ith element of $\breve{u}_1^{(m)}$. Here, ρ is the minimax concave penalty (MCP) defined as

$$\rho(t; \mu_1, a) = \mu_1 \int_0^t (1 - x/(\mu_1 a))_+ dx$$

with derivative $\dot{\rho}(t; \mu_1, a) = \mu_1(1 - t/(\mu_1 a))_+$ and $(t)_+ = \max(0, t)$, regularization parameters a and b, and with tuning parameter μ_1; see, e.g., [25]. Here we adopt the composite MCP: the outer penalty determines whether a variable is important in any of the PCs; and for an important variable, the inner penalty determines in which dataset(s) it is important. For integrative analysis under the "regression analysis + variable selection" framework, the composite MCP has been adopted and shown to be effective [11]. Here we adopt it in a significantly different context. Note that the composite MCP can be replaced by other composite and sparse group penalties which can also conduct two-level selection.

2.4. iSPCA with contrasted penalization

With the composite MCP (and other composite and sparse group penalties), there is almost no account for the relationships among datasets. Our exploratory analysis, which is partly shown in simulation below, suggests that although the approach described above outperforms the classic meta-analysis and single-dataset analysis, there is still room for improvement. For some practical data analyses, analyzing meta-data and/or comparing individual-dataset analysis results may suggest a certain degree of similarity in not just the loadings' sparsity structures but also their magnitudes/signs [3,22]. In addition, in some analyses [19], important variables with similar magnitudes/signs are of more interest, and hence it is desirable to encourage the identification of such variables. In the following, we develop two approaches to address such needs.

2.4.1. Magnitude-based contrasted penalization

When it is reasonable to expect/encourage the first PCs of the datasets to have loadings with similar magnitudes, we propose imposing the following *magnitude-based contrasted penalty* in addition to the penalty function in (1):

$$\frac{\mu_2}{2}\sum_{i=1}^{d}\sum_{1\le\ell<m\le M}(\breve{u}_{i,1}^{(m)}-\breve{u}_{i,1}^{(\ell)})^2, \tag{2}$$

where $\mu_2 > 0$ is a data-dependent tuning parameter. We refer to this approach as iSPCA$_M$, where the subscript "M" emphasizes magnitude.

This approach involves two penalties. The first is the composite MCP and has the same interpretation as described in the above section. For the ith variable, the newly added penalty (2) encourages its loadings in different datasets to have similar magnitudes. The degree of similarity is adjusted using the tuning parameter μ_2. As an alternative to the ℓ_2 penalty, we can also penalize ℓ_1 differences, which may generate estimates that are exactly equal. As we are mainly interested in similarity as opposed to exact equality, we choose the ℓ_2 penalty, which may be computationally simpler.

Shrinking the differences between parameter estimates has been considered in the literature under contexts different from the present one. Relevant approaches include the fused penalization and Laplacian penalization. The "standard" fused penalization demands a spatial structure and shrinks differences between adjacent parameters. In contrast, the proposed approach considers all (m, ℓ) pairs.

There are more recent fused penalization methods [21,23], which also impose magnitude-based contrasted penalties. However, they are under regression settings and used for the purpose of grouping. The Laplacian penalization demands a degree of adjacency measure, which is used to adjust the level of penalty and does not exist for independent datasets. A contrasted penalization approach similar to the proposed one has already been developed [19] under the "regression analysis + variable selection" framework. Note that under the present data setting, different PCs are allowed to have different sparsity structures. That is, one variable may have a nonzero loading in one PC but a zero loading in another PC. In this case, the magnitude-based penalty is still sensible: it shrinks the nonzero loading towards zero (or the other way around) and hence encourages similarity.

2.4.2. Sign-based contrasted penalization

Under certain scenarios, it can be reasonable to expect/encourage the first PCs of the M datasets to have loadings with similar signs. See, e.g., [11,19,26] for discussion. Note that, loosely speaking, similarity in signs is weaker than that in magnitudes. Here we propose imposing the following *sign-based contrasted penalty* in addition to the penalty function in (1):

$$\frac{\mu_2}{2}\sum_{i=1}^{d}\sum_{1\le\ell<m\le M}\{\text{sign}(\breve{u}_{i,1}^{(m)})-\text{sign}(\breve{u}_{i,1}^{(\ell)})\}^2,$$

where $\mu_2 > 0$ is a data-dependent tuning parameter, and $\text{sign}(t) = 0, 1, -1$ depending whether $t = 0, t > 0$ or $t < 0$.

The interpretation of this penalty is similar to that of (2). We refer to this approach as iSPCA$_S$, where the subscript "S" stands for sign. The sign-based penalty is not continuous, leading to challenges in optimization. To solve this problem, we further propose the approximated penalty

$$\frac{\mu_2}{2}\sum_{i=1}^{d}\sum_{1\le\ell<m\le M}\left(\frac{\breve{u}_{i,1}^{(m)}}{\sqrt{\breve{u}_{i,1}^{(m)2}+\tau^2}}-\frac{\breve{u}_{i,1}^{(\ell)}}{\sqrt{\breve{u}_{i,1}^{(\ell)2}+\tau^2}}\right)^2,$$

where $\tau > 0$ is a small positive constant. Penalties based on the sign function (and its approximations) have been considered in [1], but the data settings and analysis goals are quite different from the present study. To the best of our knowledge, sign-based penalties have not been well adopted in integrative analysis, especially not in the context of dimension reduction.

Remarks. Note that estimates of the PCs are unique up to a sign. To avoid sign swamps across datasets, we set the loadings with the largest absolute values to have a positive sign. For both iSPCA$_M$ and iSPCA$_S$, computation of the second and other downstream PCs can be conducted consecutively by keeping updating the data matrices [16]. In practical data analysis, the

number of PCs may also need to be determined. In single-dataset analysis, one approach is to select the number of PCs so that a predetermined cumulative percentage of explained variance (CPEV) is reached. An alternative approach is to jointly examine eigenvalues and variations of eigenvectors. Both approaches can be extended to the present setting. For example, for each dataset, we can select the number of PCs to meet the CPEV criterion.

2.5. Statistical properties

Here we establish statistical properties of iSPCA$_S$ in high-dimensional settings. With a simpler contrasted penalty, properties of iSPCA$_M$ can be established in a similar manner. We omit the details for iSPCA$_M$ here. Under iSPCA$_S$, the overall penalized objective function is:

$$L(\breve{u}_1^{(1)}, \ldots, \breve{u}_1^{(M)}) = \sum_{m=1}^{M} \frac{1}{2n_m} \|X^{(m)} - \breve{u}_1^{(m)} \tilde{v}_1^{(m)\top}\|_F^2 + \sum_{i=1}^{d} \rho\left(\sum_{m=1}^{M} \rho(|\breve{u}_{i,1}^{(m)}|; \mu_1, a); 1, b\right) + \frac{\mu_2}{2} \sum_{i=1}^{d} \sum_{1 \le \ell < m \le M} \left(\frac{\breve{u}_{i,1}^{(m)}}{\sqrt{\breve{u}_{i,1}^{(m)2} + \tau^2}} - \frac{\breve{u}_{i,1}^{(\ell)}}{\sqrt{\breve{u}_{i,1}^{(\ell)2} + \tau^2}} \right)^2, \quad \text{subject to } \|\tilde{v}_1^{(1)}\| = \cdots = \|\tilde{v}_1^{(M)}\| = 1. \tag{3}$$

For a single dataset, the concept of consistency of PCA has been proposed in [7] and others. Specifically, let $\bar{u}_i$ be a sample-based estimator for u_i for $i \in \{1, \ldots, d\}$. The direction $\bar{u}_i$ is consistent for its population counterpart u_i if $\text{Angle}(\bar{u}_i, u_i) = \arccos(|\langle \bar{u}_i, u_i \rangle|) \xrightarrow{p} 0$ as $d \to \infty$, where $\langle \cdot \rangle$ denotes the inner product of two vectors.

For the proposed analysis, we establish below that for each $m \in \{1, \ldots, M\}$, the principal component direction $\tilde{u}_1^{(m)} = \breve{u}_1^{(m)} / \|\breve{u}_1^{(m)}\|$ is consistent for $u_1^{(m)}$. The following conditions are assumed.

Condition 1. *For $m \in \{1, \ldots, M\}$, $X_1^{(m)}, \ldots, X_{n_m}^{(m)}$ are iid observations from a d-dimensional multivariate normal distribution $\mathcal{N}(0, \Sigma_d^{(m)})$. The eigenvalues of $\Sigma_d^{(m)}$ satisfy*

$$\lambda_1^{(m)} \sim d^{\alpha^{(m)}}, \quad \lambda_2^{(m)} \sim d^{\theta^{(m)}}, \quad \sum_{i=2}^{d} \lambda_i^{(m)} \sim d, \tag{4}$$

where $\theta^{(m)} \in [0, \alpha^{(m)})$ and $\alpha^{(m)} \in (0, 1]$.

Condition 2. *For $m \in \{1, \ldots, M\}$, denote $A^{(m)} = \{i \in \{1, \ldots, d\} : u_{i,1}^{(m)} \neq 0\}$ where $u_{i,1}^{(m)}$ is the ith element of $u_1^{(m)}$. Then $|A^{(m)}| = \lfloor d^{\beta^{(m)}} \rfloor$, where $|A^{(m)}|$ is the cardinality of $A^{(m)}$, $\beta^{(m)} \in [0, 1]$ is a sparsity parameter, and $\lfloor d^{\beta^{(m)}} \rfloor$ is the integer part of $d^{\beta^{(m)}}$. Assume that $\max_{i \in A^{(m)}} |u_{i,1}^{(m)}|^{-1} \sim d^{\eta^{(m)}/2}$, where $\eta^{(m)} \in [0, \alpha^{(m)})$ and $\alpha^{(m)}$ is the spike parameter.*

Condition 3. *$H_j^{(m)} \equiv \sum_{i=2}^{d} X_j^{(m)\top} u_i^{(m)} u_i^{(m)}$ satisfies the mixing condition for all $j \in \{1, \ldots, n_m\}$ and $m \in \{1, \ldots, M\}$.*

Condition 4. *$\mu_1 = o\{d^{(\min(\alpha^{(m)} - \eta^{(m)}) - \kappa)/2}\}$ where $\kappa \in (1 - \min(\alpha^{(m)}), \min(\alpha^{(m)} - \eta^{(m)}) - \max(\theta^{(m)}))$ and $\mu_2 = o(\mu_1)$.*

Condition 5. *$\sum_{i=2}^{d} \lambda_i^{(m)2} / (\sum_{i=2}^{d} \lambda_i^{(m)})^2 \to 0$ as $d \to \infty$, for all $m \in \{1, \ldots, M\}$.*

The assumed conditions are mild and not stronger than their counterparts in single-dataset PCA studies. Specifically, the normality assumption in Condition 1 is common for PCA and SPCA studies [5,15,17]. Condition (4) ensures that $\Sigma_d^{(m)}$ has a single component spike structure. In addition, the second eigenvalue $\lambda_2^{(m)}$ is bounded away from the first eigenvalue $\lambda_1^{(m)}$. The ranges of $\alpha^{(m)}$ and $\theta^{(m)}$ are the same as in [17] and other publications.

Condition 2 assumes that $u_1^{(m)}$ is sparse and that the speed of the smallest element in $u_1^{(m)}$ shrinking towards to 0 is slower than $d^{-\eta^{(m)}/2}$. Define

$$Z_j^{(m)} \equiv (z_{1,j}^{(m)}, \ldots, z_{d,j}^{(m)})^\top = (X_j^{(m)\top} u_1^{(m)}, \ldots, X_j^{(m)\top} u_d^{(m)})^\top.$$

Then, for all $j \in \{1, \ldots, n_m\}$,

$$H_j^{(m)} = (h_{1,j}^{(m)}, \ldots, h_{d,j}^{(m)})^\top = \sum_{i=2}^{d} z_{i,j}^{(m)} u_i^{(m)}.$$

To establish consistency, it is needed that the $H_j^{(m)}$s have a negligible effect on the direction vector $\tilde{u}_1^{(m)}$.

Suppose that the $H_j^{(m)}$s are iid $\mathcal{N}(0, \Delta_d^{(m)})$, where $\Delta_d^{(m)} = (r_{k\ell}^{(m)})_{d \times d}$. A sufficient condition under which their effect is negligible is the following mixing condition of [8]:

$$\forall_{k \neq \ell \in A^{(m)}} \quad |r_{k\ell}^{(m)}| \le r_{kk}^{(m)1/2} r_{\ell\ell}^{(m)1/2} \rho_{|k-\ell|}^{(m)},$$

where $\rho_t^{(m)} < 1$ for all $t > 1$ and $\rho_t^{(m)} \ln(t) \to 0$, as $t \to \infty$. Condition 4 controls the divergence speed of the tuning parameters. Condition 5 has been referred to as the ϵ_2-condition [7], and

$$\epsilon_2^{(m)} = \left(\sum_{i=2}^{d} \lambda_i^{(m)}\right)^2 \Big/ \left(d \sum_{i=2}^{d} \lambda_i^{(m)2}\right)$$

is the measure of sphericity for $\lambda_2^{(m)}, \ldots, \lambda_d^{(m)}$.

Properties of the iSPCA$_S$ estimates can be summarized as follows.

Theorem 1. *Suppose that Conditions 1–4 hold. For each $m \in \{1, \ldots, M\}$, fix $\tilde{v}_1^{(m)}$at $\hat{v}_1^{(m)}$ which is the first right singular vector of $X^{(m)}$. Then $\tilde{u}_1^{(m)} = \breve{u}_1^{(m)}/\|\breve{u}_1^{(m)}\|$ is consistent for $u_1^{(m)}$, with a convergence rate of $d^{\kappa/2}$, where $(\breve{u}_1^{(1)}, \ldots, \breve{u}_1^{(M)})$ optimize (3) with $\tilde{v}_1^{(m)} = \hat{v}_1^{(m)}$.*

Theorem 1 establishes that with properly chosen tunings, there exists a $d^{\kappa/2}$ consistent estimator for $u_1^{(m)}$ for each $m \in \{1, \ldots, M\}$ as $d \to \infty$. The rate of convergence is affected by the divergence speed of $\lambda_1^{(m)}$ and $\lambda_2^{(m)}$ as well as the speed of the smallest element of $u_1^{(m)}$ shrinking towards 0. More specifically, with a larger divergence rate of $\lambda_1^{(m)}$, a lower divergence rate of $\lambda_2^{(m)}$, and a lower convergence rate of the smallest element of $u_1^{(m)}$, the estimator has a higher convergence rate. We can also fix $\tilde{v}_1^{(m)}$ in a general way, and the consistency properties can be summarized as follows.

Theorem 2. *Suppose that Conditions 1–5 hold. Suppose that $\hat{u}_1^{0,(m)}$ is consistent for $u_1^{(m)}$, with a convergence rate of $d^{\kappa/2}$. For each $m \in \{1, \ldots, M\}$, fix $\tilde{v}_1^{(m)}$ at $X^{(m)\top}\hat{u}_1^{0,(m)}/\|X^{(m)\top}\hat{u}_1^{0,(m)}\|$. Then $\tilde{u}_{1,\text{new}}^{(m)} = \breve{u}_{1,\text{new}}^{(m)}/\|\breve{u}_{1,\text{new}}^{(m)}\|$ is consistent for $u_1^{(m)}$, with a convergence rate of $d^{\kappa/2}$, where $(\breve{u}_{1,\text{new}}^{(1)}, \ldots, \breve{u}_{1,\text{new}}^{(M)})$ optimize (3) with fixed $\tilde{v}_1^{(m)}$.*

The main difference between Theorems 1 and 2 is on how to choose $\tilde{v}_1^{(m)}$. In Theorem 1, we fix $\tilde{v}_1^{(m)}$ as the first right singular vector of $X^{(m)}$ for each $m \in \{1, \ldots, M\}$, while in Theorem 2, we obtain $\tilde{v}_1^{(m)}$ based on a consistent estimator of $u_1^{(m)}$. Theorems 1–2 suggest that we can take an iterative algorithm to optimize (3), and the scale vector $\tilde{u}_1^{(m)}$ of $\breve{u}_1^{(m)}$ is consistent for $u_1^{(m)}$ at every updating step.

2.6. Computation

Here we consider iSPCA$_S$. The computational algorithm for iSPCA$_M$ is described in the Appendix. The proposed algorithm is summarized in Algorithm 1. It is iterative and optimizes over $\breve{u}_1^{(m)}$ for fixed $\tilde{v}_1^{(m)}$ for each $m \in \{1, \ldots, M\}$. Then it optimizes over $\tilde{v}_1^{(m)}$ for fixed $\breve{u}_1^{(m)}$ for each $m \in \{1, \ldots, M\}$. The algorithm is repeated until convergence.

Algorithm 1: Computational Algorithm for iSPCA$_S$

1. Initialize: For each $m \in \{1, \ldots, M\}$, apply the standard SVD to $X^{(m)}$, and obtain the best rank-1 approximation of $X^{(m)}$ as $\hat{\omega}^{(m)}\hat{u}_1^{(m)}\hat{v}_1^{(m)\top}$, where $\hat{u}_1^{(m)}$ and $\hat{v}_1^{(m)}$ are unit-norm vectors and $\hat{\omega}^{(m)}$ is the first singular value of $X^{(m)}$. Set $\breve{u}_{1,\text{old}}^{(m)} = \hat{\omega}^{(m)}\hat{u}_1^{(m)}$ and $\tilde{v}_{1,\text{old}}^{(m)} = \hat{v}_1^{(m)}$ for each $m \in \{1, \ldots, M\}$.
2. Update:
   (a) $\breve{\mathbf{u}}_{1,\text{new}} \equiv (\breve{u}_{1,\text{new}}^{(1)}, \ldots, \breve{u}_{1,\text{new}}^{(M)}) = f(\tilde{v}_{1,\text{old}}^{(1)}, \ldots, \tilde{v}_{1,\text{old}}^{(M)})$.
   (b) $\tilde{v}_{1,\text{new}}^{(m)} = X^{(m)\top}\breve{u}_{1,\text{new}}^{(m)}/\|X^{(m)\top}\breve{u}_{1,\text{new}}^{(m)}\|$ for each $m \in \{1, \ldots, M\}$.
3. Repeat Step 2 while replacing $\breve{u}_{1,\text{old}}^{(m)}$ and $\tilde{v}_{1,\text{old}}^{(m)}$ by $\breve{u}_{1,\text{new}}^{(m)}$ and $\tilde{v}_{1,\text{new}}^{(m)}$ respectively until convergence. In our numerical study, we used the ℓ_2 norm of the difference between two consecutive estimates smaller than a prefixed threshold as the criterion for convergence.
4. Standardize the final $\breve{u}_{1,\text{new}}^{(m)}$ as $\tilde{u}_1^{(m)} = \breve{u}_{1,\text{new}}^{(m)}/\|\breve{u}_{1,\text{new}}^{(m)}\|$ for each $m \in \{1, \ldots, M\}$.

In Algorithm 1 , the key is Step 2(a). With fixed $\tilde{v}_1^{(1)}, \ldots, \tilde{v}_1^{(M)}$, the function f is defined via the following:

$$\breve{\mathbf{u}}_1 = f(\tilde{v}_1^{(1)}, \ldots, \tilde{v}_1^{(M)}) = \underset{\breve{u}_1^{(m)}, m\in\{1,\ldots,M\}}{\operatorname{argmin}} \; L(\breve{u}_1^{(1)}, \ldots, \breve{u}_1^{(M)}),$$

where L is the objective function in (3). As

$$\sum_{m=1}^{M}\frac{1}{2n_m}\|X^{(m)}-\breve{u}_1^{(m)}\tilde{v}_1^{(m)\top}\|_F^2=\sum_{i=1}^{d}\sum_{m=1}^{M}\frac{1}{2n_m}\sum_{j=1}^{n_m}(x_{ij}^{(m)}-\breve{u}_{i,1}^{(m)}\tilde{v}_{j,1}^{(m)})^2,$$

the lack-of-fit measure is separable. In addition, the penalty function is also separable for the rows of $\breve{\mathbf{u}}_1$. Therefore, we can optimize over each individual row of $\breve{\mathbf{u}}_1$. Suppose that we have an initial estimate $\{\breve{u}_{i,1}^{(1)0},\ldots,\breve{u}_{i,1}^{(M)0}\}$ which is close to the minimizer and which can be obtained from the standard SPCA. For $\breve{u}_{i,1}^{(m)}$,

$$L(\breve{u}_{i,1}^{(m)})\approx\frac{1}{2n_m}\sum_{j=1}^{n_m}(x_{ij}^{(m)}-\breve{u}_{i,1}^{(m)}\tilde{v}_{j,1}^{(m)})^2+\mu_{im}(|\breve{u}_{i,1}^{(m)}|-|\breve{u}_{i,1}^{(m)0}|)+\frac{\mu_2}{2}\sum_{\ell\neq m}\left\{\frac{\breve{u}_{i,1}^{(m)}}{\sqrt{(\breve{u}_{i,1}^{(m)0})^2+\tau^2}}-\frac{\breve{u}_{i,1}^{(\ell)0}}{\sqrt{(\breve{u}_{i,1}^{(\ell)0})^2+\tau^2}}\right\}^2,$$

where $\mu_{im}=\dot{\rho}(\sum_{\ell=1}^{M}\rho(|\breve{u}_{i,1}^{(\ell)0}|;\mu_1,a),1,b)\dot{\rho}(|\breve{u}_{i,1}^{(m)0}|;\mu_1,a)$. Then

$$\breve{u}_{i,1}^{(m)}=n_m[1+\mu_2 n_m(M-1)/\{(\breve{u}_{i,1}^{(m)0})^2+\tau^2\}]^{-1}S(Z_i^{(m)0},\mu_{im}),$$

where $S(z,t)=\text{sign}(z)(|z|-t)_+$, and

$$Z_i^{(m)0}=\frac{1}{n_m}\sum_{j=1}^{n_m}x_{ij}^{(m)}\tilde{v}_{j,1}^{(m)}+\frac{\mu_2}{\sqrt{(\breve{u}_{i,1}^{(m)0})^2+\tau^2}}\sum_{\ell\neq m}\frac{\breve{u}_{i,1}^{(\ell)0}}{\sqrt{(\breve{u}_{i,1}^{(\ell)0})^2+\tau^2}}.$$

Thus, in Step 2 (a), we can solve for $\breve{\mathbf{u}}_{1,\text{new}}$ as follows. For the ith row of $\breve{\mathbf{u}}_{1,\text{new}}$, with $i\in\{1,\ldots,d\}$,

(1) let $\{\breve{u}_{i,1}^{(1)0},\ldots,\breve{u}_{i,1}^{(M)0}\}=\{\breve{u}_{i,1,old}^{(1)},\ldots,\breve{u}_{i,1,old}^{(M)}\}$, and fix $\{\tilde{v}_1^{(1)},\ldots,\tilde{v}_1^{(M)}\}=\{\tilde{v}_{1,\text{old}}^{(1)},\ldots,\tilde{v}_{1,\text{old}}^{(M)}\}$;
(2) for each $m\in\{1,\ldots,M\}$,
  (a) update $\mu_{im}=\dot{\rho}\left(\sum_{\ell=1}^{M}\rho(|\breve{u}_{i,1}^{(\ell)0}|;\mu_1,a),1,b\right)\dot{\rho}(|\breve{u}_{i,1}^{(m)0}|;\mu_1,a)$;
  (b) update $\breve{u}_{i,1}^{(m)0}=n_m[1+\mu_2 n_m(M-1)/\{(\breve{u}_{i,1}^{(m)0})^2+\tau^2\}]^{-1}S(Z_i^{(m)0},\mu_{im})$;
(3) return $(\breve{u}_{i,1}^{(1)0},\ldots,\breve{u}_{i,1}^{(M)0})$.

With the proposed algorithm, computational cost increases linearly with d. To reduce computing time, estimation for different i can be performed in a parallel manner. In each iteration, we only conduct Step 2 once as opposed to repeating until convergence, which not only increases computational complexity but also decreases convergence speed with the linkage between $\breve{u}_1^{(m)}$ and $\tilde{v}_1^{(m)}$. Convergence of the proposed algorithm follows along the same lines as that for single-dataset SPCA, which has been studied in the literature [17]. In all of our numerical studies, convergence is achieved within a small to moderate number of iterations. With simple updates and fast convergence, the proposed algorithm is computationally much affordable. For example, for one simulation replicate with $d=500$ (more details described in the next section), the analysis takes 49.8 s on a laptop with standard configurations.

2.6.1. Tuning parameter selection

iSPCA$_S$ involves tuning parameters μ_1,μ_2 and regularization parameters a,b. For a in MCP, published studies [11,25] suggest examining a small number of values or fixing its value. In our case, we found that the results are not sensitive to a. We set $a=6$ as suggested in the literature. For b, we set $b=1/2Ma\mu_1^2$, which has been suggested in published composite MCP studies, with the linkage between the inner and outer penalties. We use cross-validation to choose μ_1 and μ_2. Furthermore, iSPCA$_S$ involves τ, with a smaller value leading to a better approximation but less stable computation. Closely examining the methodological and theoretical developments suggests that the proposed approach does not demand a very accurate approximation of the sign function. As long as τ is not too big and the approximation can differentiate parameters with different signs, the proposed approach is valid. In our study, we fixed the value of $\tau^2=0.5$, which leads to satisfactory results. Other values may also need to be considered in practice.

3. Numerical study

3.1. Simulation

We conduct simulation to gauge performance of the proposed approach and compare with alternatives. We set $M=4$, $n_m=25$, and consider $d\in\{500,1000\}$. To generate the data matrices, we first need to determine the population covariance matrices. Following published single-dataset SPCA studies [17], we consider the single spike covariance structure with

Table 1
Simulation: combinations of α and β values.

	Case 1	Case 2	Case 3
$(\beta^{(1)}, \beta^{(2)}, \beta^{(3)}, \beta^{(4)})$	$(\alpha^{(1)}, \alpha^{(2)}, \alpha^{(3)}, \alpha^{(4)})$,	$(\alpha^{(1)}, \alpha^{(2)}, \alpha^{(3)}, \alpha^{(4)})$	$(\alpha^{(1)}, \alpha^{(2)}, \alpha^{(3)}, \alpha^{(4)})$
(0.3, 0.3, 0.3, 0.3)	(0.4, 0.4, 0.4, 0.4)	(0.4, 0.5, 0.6, 0.7)	(0.3, 0.4, 0.4, 0.5)
(0.5, 0.5, 0.5, 0.5)	(0.6, 0.6, 0.6, 0.6)	(0.6, 0.6, 1.0, 1.5)	(0.4, 0.6, 0.6, 0.7)
(0.8, 0.8, 0.8, 0.8)	(1.0, 1.0, 1.0, 1.0)	(0.9, 0.9, 1.5, 1.5)	(0.7, 0.9, 0.9, 1.0)

$\lambda_1^{(m)} = d^{\alpha^{(m)}}$, $\lambda_2^{(m)} = \cdots = \lambda_d^{(m)} = 1$, for all $m \in \{1, \ldots, M\}$. We construct the covariance matrices $\Sigma_d^{(m)}$s using the eigen-decomposing expression

$$\Sigma_d^{(m)} = \sum_{i=1}^{d} \lambda_i^{(m)} u_i^{(m)} u_i^{(m)\top},$$

where $u_i^{(m)}$ is the ith eigenvector. Set the number of nonzero entries in $u_1^{(m)}$ as $\lfloor d^{\beta^{(m)}} \rfloor$. To generate the orthonormal eigenvectors of $\Sigma_d^{(m)}$, we first form a full rank matrix $U_d^{(m)*} = [u_1^{(m)*}, \ldots, u_d^{(m)*}]$. Then, we apply the Gram–Schmidt orthogonalization to $U_d^{(m)*}$ to obtain the orthogonal matrix $U_d^{(m)} = [u_1^{(m)}, \ldots, u_d^{(m)}]$. For $u_1^{(m)*}$, we consider the following four scenarios:

Scenario I: $u_1^{(m)*} = (\overbrace{1, \ldots, 1}^{\lfloor d^{\beta^{(m)}} \rfloor}, 0, \ldots, 0)^\top$ for each $m \in \{1, \ldots, M\}$. That is, the $u_1^{(m)*}$s have the same sparsity structure as well as the same values.

Scenario II: First generate $u_{i,1}^{(m)*} \sim \mathcal{N}((\lfloor d^{\beta^{(m)}} \rfloor + 1 - i)^{1.5}, (\lfloor d^{\beta^{(m)}} \rfloor + 1 - i/4)^2)$ for each $i \in \{1, \ldots, \lfloor d^{\beta^{(m)}} \rfloor\}$, and then randomly permute their values. Set $u_{i,1}^{(m)*} = 0$ for each $i \in \{\lfloor d^{\beta^{(m)}} \rfloor + 1, \ldots, d\}$. That is, the $u_1^{(m)*}$s have the same sparsity structure but not the same nonzero values.

Scenario III: $u_{i,1}^{(m)*} \sim \mathcal{N}(3, 0.2^2)$ for each $i \in \{1 + (m-1)\lfloor d^{\beta^{(m)}}/4 \rfloor, \ldots, m\lfloor d^{\beta^{(m)}}/4 \rfloor\}$; $u_{i,1}^{(m)*} \sim \mathcal{U}(0.5, 1)$ for each $i \in \{M\lfloor d^{\beta^{(m)}}/4 \rfloor + 1, \ldots, M\lfloor d^{\beta^{(m)}}/4 \rfloor + \lfloor d^{\beta^{(m)}} \rfloor - \lfloor d^{\beta^{(m)}}/4 \rfloor\}$; $u_{i,1}^{(m)*} = 0$, otherwise. That is, the $u_1^{(m)*}$s have partially overlapping sparsity structures.

Scenario IV: For randomly chosen $\{m_1, \ldots, m_{\lfloor d^{\beta^{(m)}} \rfloor}\} \subset \{1 \ldots, d\}$, $u_{m_i,1}^{(m)*} \sim \mathcal{N}(0.8, 0.7^2)$. The other components are zero. That is, the $u_1^{(m)*}$s have random sparsity structures with random overlappings.

The four scenarios comprehensively cover different degrees of overlapping in sparsity structure and similarity in nonzero values. The components of $u_i^{(m)*}$ for $i \in \{2, \ldots, d\}$ and $m \in \{1, \ldots, M\}$ are randomly drawn from $\mathcal{U}(1, 2)$.

We consider multiple combinations for the spike and sparsity pairs $(\alpha^{(1)}, \beta^{(1)}), \ldots, (\alpha^{(M)}, \beta^{(M)})$. Published single-dataset SPCA studies suggest the critical role of α and β. Specifically, when $\alpha \in (0, 1]$ and $\alpha \leq \beta$, the SPCA estimation is not consistent [17]. As shown in Table 1, we consider a total of nine pairs, which can be classified into three cases. Under Case 1, the $\alpha^{(m)}$s are equal across the M datasets. Under Case 2, $\alpha^{(m)} > \beta^{(m)}$ for all of the M datasets. Under Case 3, $\alpha^{(1)} \leq \beta^{(1)}$. For each simulation setting, we simulate 200 replicates.

Beyond iSPCA (the approach described in Section 2.2 without contrasted penalization), iSPCA$_M$, and iSPCA$_S$, we also consider mPCA, mSPCA, and sSPCA. mPCA and mSPCA conduct PCA and SPCA for each dataset separately, and then summary statistics (the sets of variables with nonzero loadings) are combined across datasets. Here "m" stands for meta-analysis. Under sSPCA, all M datasets are stacked together and then analyzed using SPCA. Here "s" stands for stacking. In the literature, multi-datasets dimension reduction analysis is very limited. The three alternatives are the most "straightforward" and relevant. To compare different approaches, we consider the following measures: angle (the acute angle between the estimated and true first eigenvectors), TPR (true positive rate), and FDR (false discovery rate). Among them, angle quantifies estimation accuracy, while TPR and FDR quantify identification accuracy.

The estimation results are summarized in Table 2 for Case 1 and in Tables 5–6 in the Appendix for Cases 2–3. It is observed that the proposed integrative analysis has competitive performance. Specifically, mPCA, which generates dense estimates, is inferior. Under Scenarios I, II, and III, which have overlapping sparsity structures, integrative analysis outperforms mSPCA. sSPCA excels under Scenario I, which is as expected, but has inferior performance under Scenarios II, III, and IV. Under Scenario I, where variables have the same loadings across datasets, iSPCA$_M$ has the best performance, as expected. Under Scenarios II and III, iSPCA$_S$, which is less stringent, outperforms the others. Under Scenario IV, which does not favor multi-datasets analysis, the proposed integrative analysis still has performance comparable to meta-analysis. Under Case 3, which may have theoretically inconsistent estimates, the proposed analysis still has reasonable performance. For mSPCA and the three integrative analysis methods, the identification results are summarized in Fig. 1. For sSPCA, the identification results are summarized in Table 7 in the Appendix (as the FDR values are dramatically larger than those in Fig. 1). Under Scenario I, all methods have TPR values close to 1 and FDR values close to 0 (details omitted). Fig. 1 and Table 7 document the competitive identification performance of the proposed analysis. Overall, iSPCA$_S$ has the most favorable identification performance, and the other two integrative analysis methods are competitive compared to the meta-analysis methods and sSPCA.

Table 2
Simulation Case 1. In each cell: median of the average angles across the M datasets between the estimated and true loading vectors (median absolute deviation/0.6745).

β	$d = 500$						$d = 1000$					
	mPCA	mSPCA	sSPCA	iSPCA	iSPCA$_M$	iSPCA$_S$	mPCA	mSPCA	sSPCA	iSPCA	iSPCA$_M$	iSPCA$_S$
Scenario I												
0.3	58.22	10.29	3.46	8.18	3.79	7.98	63.41	9.75	3.77	7.88	3.98	7.75
	(4.74)	(6.01)	(1.38)	(2.05)	(1.26)	(1.75)	(3.40)	(4.67)	(1.26)	(1.08)	(1.14)	(0.88)
0.5	35.49	8.60	3.83	8.22	3.79	8.10	40.45	8.51	4.12	8.26	4.18	8.14
	(2.44)	(1.32)	(0.52)	(1.13)	(0.37)	(1.11)	(2.56)	(1.47)	(0.51)	(0.74)	(0.52)	(0.67)
0.8	11.92	6.47	3.15	6.47	3.18	6.40	11.35	5.73	2.80	5.73	2.81	5.71
	(0.88)	(0.38)	(0.17)	(0.39)	(0.22)	(0.38)	(0.66)	(0.29)	(0.10)	(0.29)	(0.11)	(0.28)
Scenario II												
0.3	58.27	14.55	29.74	13.96	11.14	8.22	63.00	13.89	29.82	12.53	10.71	7.52
	(3.53)	(3.54)	(2.16)	(4.05)	(3.02)	(1.90)	(2.45)	(3.30)	(3.35)	(1.74)	(2.22)	(1.14)
0.5	36.24	12.63	31.47	12.26	11.04	8.78	39.85	11.97	31.71	11.63	10.81	8.55
	(2.49)	(1.99)	(1.25)	(2.29)	(1.52)	(1.03)	(1.66)	(1.04)	(1.30)	(0.96)	(0.74)	(0.62)
0.8	11.74	8.00	31.85	7.76	7.41	6.41	12.02	7.69	32.08	7.35	7.11	6.10
	(0.82)	(0.57)	(0.40)	(0.60)	(0.57)	(0.63)	(0.68)	(0.41)	(0.31)	(0.48)	(0.53)	(0.25)
Scenario III												
0.3	60.32	22.24	54.17	19.94	15.04	8.85	62.94	20.61	50.43	15.12	11.39	7.71
	(3.55)	(2.36)	(3.48)	(4.91)	(4.94)	(2.43)	(3.27)	(3.44)	(2.66)	(4.44)	(3.12)	(1.18)
0.5	36.30	17.74	56.69	15.50	13.09	8.92	38.93	16.66	55.53	14.03	12.60	8.69
	(2.00)	(2.08)	(2.12)	(2.05)	(1.49)	(1.11)	(2.36)	(2.02)	(2.83)	(2.43)	(2.13)	(1.07)
0.8	11.59	8.60	55.84	7.37	7.12	6.35	11.75	8.13	55.68	7.08	6.81	6.04
	(1.04)	(1.21)	(3.18)	(1.01)	(0.84)	(0.41)	(0.45)	(0.75)	(2.33)	(0.50)	(0.46)	(0.41)
Scenario IV												
0.3	59.23	15.38	75.81	17.01	15.88	15.23	62.64	15.33	75.84	15.26	14.84	14.40
	(4.55)	(3.88)	(0.61)	(4.17)	(4.40)	(4.13)	(2.97)	(6.08)	(0.34)	(5.48)	(4.98)	(4.85)
0.5	37.08	13.41	74.44	14.17	13.88	13.68	38.99	12.50	74.99	13.13	12.49	12.46
	(1.49)	(1.24)	(1.62)	(2.05)	(1.70)	(1.78)	(1.49)	(1.38)	(0.77)	(1.06)	(1.35)	(1.06)
0.8	11.51	7.83	55.24	8.04	7.98	7.89	11.85	7.60	56.97	7.80	7.72	7.65
	(0.80)	(0.73)	(2.05)	(0.80)	(0.80)	(0.80)	(0.68)	(0.58)	(1.74)	(0.56)	(0.58)	(0.64)

3.2. Data analysis

Here we analyze cancer gene expression studies, which often have high-dimensional measurements on a small to moderate number of samples. In such studies, PCA (and other dimension reduction analysis) has been extensively conducted and has revealed insightful data structures as well as served as the basis for other analyses such as clustering and regression [13]. The analyzed datasets are obtained from GEO (Gene Expression Omnibus). Two sets of analysis are conducted.

3.2.1. Set I: Analysis of breast cancer datasets

The three datasets have GEO IDs GSE9574, GSE21947, and GSE5364, respectively. We refer to the original publications and GEO website for more information on these data and omit the details here except to say that the sample sizes are 29, 30, and 196, respectively. The three datasets have been jointly analyzed in published studies [20]. For each dataset, we conduct gene expression normalization (without variance standardization) and imputation of missing values. Genes are matched across datasets using their Unique IDs. In total, 20,995 gene expressions are measured in all three datasets. As gene expressions with higher variations are usually more interesting, in each dataset, we identify the 400 genes with the largest variances. Across the three datasets, 1028 gene expressions are selected for downstream analysis.

We apply the proposed methods and their sparse alternatives. The analysis results are summarized in Tables 3–4. More detailed estimation results are available from the authors. It can be seen from both tables that different methods lead to different estimation and identification results. Table 4 suggests that for this specific set of analysis, results of iSPCA and iSPCA$_S$ are the closest. The three datasets come from independent studies conducted by different research groups under different protocols. iSPCA$_M$ assumes similar magnitudes, which may be too stringent. The proposed integrative analysis methods with contrasted penalties identify more consistent nonzero loadings across datasets, which may lead to more focused downstream analysis.

To complement the estimation and identification analysis, we also evaluate the stability of analysis by computing the observed occurrence index (OOI) [4]. Briefly, we randomly select 75% samples from each dataset and then conduct integrative analysis. This process is repeated 100 times. For each gene identified (as having nonzero loadings) using the whole data, we compute its probability of being identified out of the 100 resamplings; this probability has been referred to as the OOI. For the proposed integrative analysis, the median OOI values are 0.91 (iSPCA), 0.94 (iSPCA$_M$), and 0.91 (iSPCA$_S$), respectively. The satisfactory stability provides additional support for the validity of the analysis.

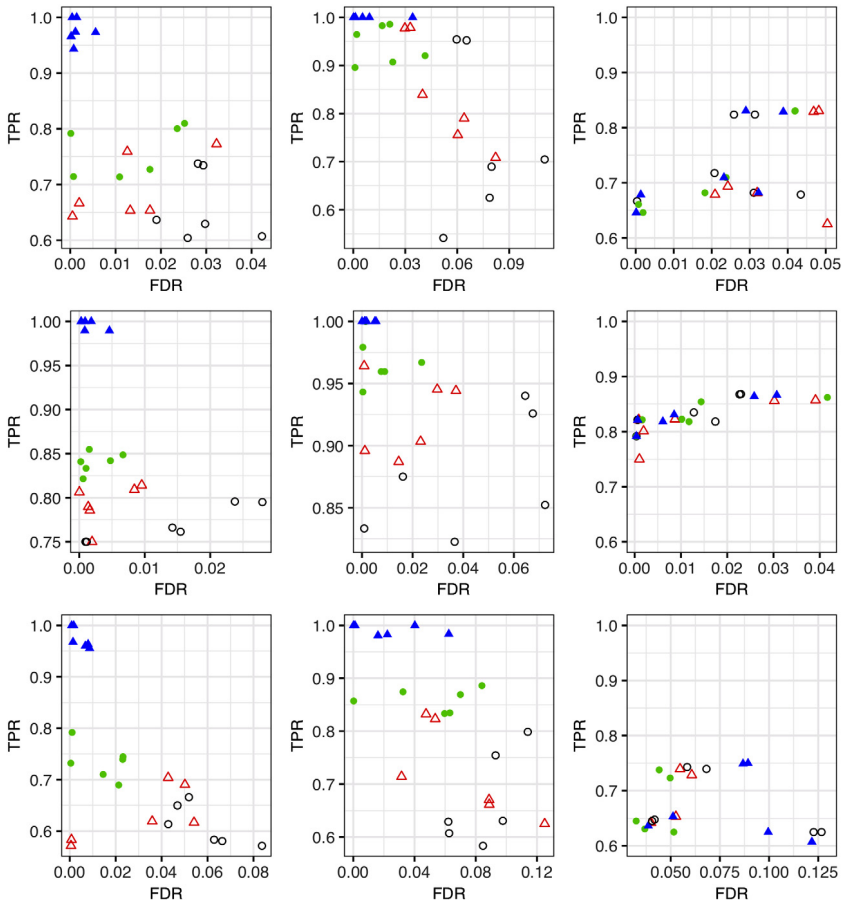

Fig. 1. Simulation: summary of identification results. Rows 1–3 correspond to Cases 1–3; Columns 1–3 correspond to Scenarios II–IV. Filled triangles, dots, hollow triangles, and circles correspond to iSPCA_S, iSPCA_M, iSPCA, and mSPCA, respectively. In each panel, there are six points for each method, corresponding to three values of $\beta^{(m)}$ and two values of d.

Table 3
Data analysis: Number of genes identified as having nonzero loadings and overlap between two datasets.

	Dataset	sSPCA	mSPCA	iSPCA	iSPCA_M	iSPCA_S
Set I						
Number	1	177	41	44	81	44
	2	177	63	77	81	75
	3	177	166	183	81	174
Overlap	1,2	177	28	38	81	38
	1,3	177	24	29	81	29
	2,3	177	44	54	81	52
Set II						
Number	1	627	53	101	101	99
	2	627	452	153	153	153
	3	627	746	322	322	322
	4	627	535	576	576	589
Overlap	1,2	627	52	84	84	82
	1,3	627	41	62	62	58
	1,4	627	35	76	76	74
	2,3	627	347	97	97	95
	2,4	627	263	111	111	113
	3,4	627	454	263	263	264

Table 4
Data analysis: angle between PC loadings estimated by different methods.

	sSPCA	mSPCA	iSPCA	iSPCA$_M$	iSPCA$_S$	sSPCA	mSPCA	iSPCA	iSPCA$_M$	iSPCA$_S$
	Set I									
GSE9574						GSE2194				
sSPCA	0.00	42.49	42.27	21.99	42.47	0.00	27.72	27.67	21.99	27.67
mSPCA		0.00	0.89	22.82	0.88		0.00	1.40	9.69	1.30
iSPCA			0.00	22.77	0.03			0.00	9.62	0.50
iSCPA$_M$				0.00	22.77				0.00	9.60
iSPCA$_S$					0.00					0.00
GSE5364										
sSPCA	0.00	5.22	5.22	21.97	5.21					
mSPCA		0.00	0.61	27.08	0.42					
iSPCA			0.00	27.08	0.44					
iSCPA$_M$				0.00	27.07					
iSPCA$_S$					0.00					
	Set II									
GSE9574						GSE16515				
sSPCA	0.00	89.11	89.09	89.09	89.10	0.00	16.87	16.89	16.89	16.89
mSPCA		0.00	2.23	2.23	2.17		0.00	0.53	0.53	0.53
iSPCA			0.00	0.00	0.51			0.00	0.00	0.13
iSCPA$_M$				0.00	0.51				0.00	0.13
iSPCA$_S$					0.00					0.00
GSE21947						GSE19650				
sSPCA	0.00	89.30	89.31	89.31	89.31	0.00	0.36	0.35	0.35	0.35
mSPCA		0.00	1.22	1.22	1.23		0.00	0.06	0.06	0.07
iSPCA			0.00	0.00	0.31			0.00	0.00	0.03
iSCPA$_M$				0.00	0.31				0.00	0.03
iSPCA$_S$					0.00					0.00

Table 5
Simulation Case 2. In each cell: median of the average angles across the M datasets between the estimated and true loading vectors (median absolute deviation/0.6745).

β	$d = 500$						$d = 1000$					
	mPCA	mSPCA	sSPCA	iSPCA	iSPCA$_M$	iSPCA$_S$	mPCA	mSPCA	sSPCA	iSPCA	iSPCA$_M$	iSPCA$_S$
Scenario I												
0.3	43.19	5.41	1.90	5.40	2.21	5.07	46.83	5.03	1.94	5.03	1.98	4.29
	(2.97)	(1.73)	(0.58)	(0.91)	(0.87)	(1.16)	(2.07)	(1.16)	(0.51)	(1.07)	(0.58)	(1.31)
0.5	25.98	5.79	0.49	5.80	1.02	2.21	28.57	5.97	0.36	5.86	0.95	2.17
	(1.60)	(1.05)	(0.14)	(0.95)	(0.31)	(0.19)	(2.33)	(1.58)	(0.06)	(1.22)	(0.25)	(0.15)
0.8	11.43	6.13	0.92	6.14	1.26	2.12	11.20	5.76	0.71	5.69	1.01	2.01
	(0.86)	(0.51)	(0.10)	(0.46)	(0.13)	(0.09)	(0.82)	(0.40)	(0.08)	(0.30)	(0.13)	(0.07)
Scenario II												
0.3	43.11	8.62	32.39	7.60	6.97	5.41	47.34	8.30	32.01	7.19	6.48	5.30
	(1.99)	(2.33)	(4.11)	(2.11)	(1.47)	(1.07)	(1.44)	(2.42)	(4.56)	(1.44)	(1.40)	(1.09)
0.5	26.46	9.39	39.95	8.19	8.43	5.82	28.95	8.82	42.26	7.62	8.09	5.81
	(2.43)	(1.33)	(2.56)	(1.43)	(1.09)	(0.83)	(2.12)	(1.24)	(1.89)	(1.08)	(1.01)	(0.79)
0.8	11.29	8.01	36.31	7.43	7.23	6.04	11.72	7.61	36.64	7.12	6.98	5.81
	(1.10)	(0.72)	(0.76)	(0.90)	(0.58)	(0.50)	(1.37)	(1.04)	(0.84)	(0.93)	(0.84)	(0.91)
Scenario III												
0.3	44.78	13.61	59.51	10.26	7.60	5.72	47.43	11.65	56.15	7.48	5.69	5.03
	(3.29)	(2.92)	(3.29)	(4.34)	(3.00)	(1.42)	(2.18)	(3.14)	(3.41)	(2.61)	(1.87)	(0.95)
0.5	26.04	12.68	67.47	9.89	8.84	6.13	28.24	12.47	67.74	9.87	8.27	6.19
	(1.36)	(1.29)	(0.39)	(1.30)	(1.66)	(0.87)	(2.34)	(2.55)	(0.28)	(2.10)	(1.72)	(0.89)
0.8	11.21	9.22	62.96	8.05	7.78	6.16	11.88	9.57	63.15	8.28	8.02	6.21
	(1.17)	(1.11)	(2.13)	(1.22)	(0.90)	(0.71)	(0.77)	(0.79)	(1.83)	(0.68)	(0.60)	(0.29)
Scenario IV												
0.3	43.66	8.45	75.55	10.16	8.87	8.72	47.13	7.92	75.54	8.69	7.78	7.40
	(2.66)	(3.00)	(0.06)	(3.99)	(3.28)	(3.05)	(2.82)	(2.32)	(0.02)	(2.20)	(2.49)	(2.36)
0.5	25.65	9.06	74.52	9.61	9.44	9.24	27.94	8.90	74.71	9.30	9.32	8.88
	(1.63)	(1.12)	(0.76)	(1.86)	(1.30)	(1.41)	(1.02)	(1.15)	(0.60)	(1.57)	(1.14)	(1.40)
0.8	10.82	7.63	63.06	7.81	7.82	7.67	11.58	7.90	64.57	8.06	8.26	7.88
	(0.90)	(0.74)	(2.89)	(0.77)	(0.62)	(0.79)	(0.96)	(0.72)	(2.81)	(0.84)	(0.84)	(0.64)

Table 6
Simulation Case 3. In each cell: median of the average angles across the M datasets between the estimated and true loading vectors (median absolute deviation/0.6745).

β	$d = 500$						$d = 1000$					
	mPCA	mSPCA	sSPCA	iSPCA	iSPCA$_M$	iSPCA$_S$	mPCA	mSPCA	sSPCA	iSPCA	iSPCA$_M$	iSPCA$_S$
Scenario I												
0.3	60.22	28.63	3.88	14.69	4.12	8.52	64.36	26.03	3.58	9.06	3.56	7.85
	(3.51)	(13.65)	(1.36)	(9.73)	(1.43)	(1.50)	(3.29)	(11.09)	(1.16)	(2.72)	(0.85)	(1.04)
0.5	40.48	19.76	4.02	13.23	4.25	9.20	44.62	22.55	3.88	15.10	4.32	9.46
	(2.88)	(4.78)	(0.93)	(3.89)	(0.91)	(0.97)	(2.66)	(7.75)	(0.53)	(3.90)	(0.57)	(0.74)
0.8	18.92	13.81	4.23	12.15	4.44	9.29	18.94	13.18	3.91	11.66	4.20	8.84
	(1.36)	(2.23)	(0.32)	(1.76)	(0.24)	(0.33)	(1.16)	(2.27)	(0.38)	(2.13)	(0.35)	(0.44)
Scenario II												
0.3	59.67	18.98	30.98	17.73	13.13	8.67	64.49	23.22	31.87	16.16	14.21	7.89
	(3.43)	(7.73)	(3.15)	(6.16)	(3.98)	(1.94)	(2.25)	(9.43)	(4.08)	(3.43)	(4.07)	(1.70)
0.5	40.62	17.05	33.10	16.12	13.44	10.86	45.07	19.39	34.08	17.68	15.17	11.24
	(2.68)	(3.70)	(1.35)	(3.15)	(1.99)	(2.03)	(2.38)	(4.80)	(1.67)	(2.93)	(1.88)	(1.17)
0.8	18.88	13.64	33.60	13.03	12.31	10.55	20.00	14.04	34.26	13.32	12.38	10.80
	(1.32)	(0.74)	(1.31)	(1.00)	(0.92)	(0.74)	(1.48)	(1.68)	(1.09)	(1.42)	(1.30)	(0.80)
Scenario III												
0.3	62.09	23.32	58.88	21.55	18.73	10.09	64.34	20.78	54.57	19.43	15.68	8.73
	(4.71)	(5.94)	(4.61)	(4.93)	(5.58)	(4.33)	(2.43)	(3.40)	(4.48)	(4.28)	(3.92)	(2.56)
0.5	40.20	19.54	61.85	19.19	16.76	11.88	43.74	18.45	65.21	19.35	17.29	12.26
	(1.22)	(2.39)	(3.84)	(1.98)	(1.32)	(2.01)	(2.39)	(1.67)	(2.62)	(2.03)	(2.47)	(2.78)
0.8	18.48	15.78	63.80	14.65	14.36	10.98	20.38	16.31	64.77	15.49	15.01	11.88
	(1.38)	(0.99)	(3.28)	(1.21)	(1.24)	(1.16)	(1.33)	(0.68)	(3.18)	(1.39)	(1.33)	(0.91)
Scenario IV												
0.3	60.63	22.69	75.65	23.76	22.83	21.74	63.06	24.94	75.62	27.01	24.58	21.51
	(2.46)	(9.34)	(0.24)	(8.53)	(8.29)	(9.41)	(3.52)	(16.64)	(0.12)	(17.58)	(14.32)	(9.33)
0.5	40.20	17.39	74.69	18.10	17.80	17.66	43.66	18.57	74.83	19.90	19.17	18.59
	(2.24)	(2.41)	(0.76)	(2.73)	(1.55)	(2.04)	(1.85)	(3.21)	(0.86)	(3.65)	(2.83)	(2.92)
0.8	18.35	14.26	63.75	14.43	14.44	14.12	19.77	14.87	66.91	15.05	14.91	14.66
	(1.11)	(1.57)	(3.85)	(1.30)	(1.17)	(1.23)	(1.30)	(1.60)	(2.98)	(1.80)	(1.68)	(1.42)

3.2.2. Set II: Analysis of breast and pancreatic cancer datasets

Cancer types are intrinsically different. However, more and more studies confirm that seemingly different cancers can share common genetic ground. Thus, the joint analysis of data on different cancer types has been conducted in a series of studies and led to insightful findings [9,11,14]. Here we conduct the joint analysis of data on breast cancer (GEO ID GSE9574 and GSE21947) and pancreatic cancer (GEO ID GSE16515 and GSE19650). The sample sizes are 29, 30, 52, and 22, respectively. Pre-processing is conducted in a similar manner as described above. The four datasets share 20,985 gene expression measurements. A total of 1634 genes are selected for downstream analysis. With data on two different cancer types, the degree of similarity across datasets may be lower than in the previous analysis.

The analysis results are also summarized in Tables 3–4. Some findings are similar to those described above: different methods generate different results; by promoting similarity, iSPCA$_M$ and iSPCA$_S$ generate more consistent findings across datasets. In this set of analyses, two different cancer types are considered. It is observed that they share common genes with nonzero loadings. However, as expected, such overlap is lower than that for datasets on the same cancer type. We also conduct stability evaluation in the same way as described above. The median OOI values are 0.89 (iSPCA), 0.90 (iSPCA$_M$), and 0.89 (iSPCA$_S$), respectively. Again, satisfactory stability is observed.

4. Discussion

In the analysis of high-dimensional data, PCA and other dimension reduction techniques have been extensively applied. In this study, built on the SPCA technique, we have developed the iSPCA approach for the integrative analysis of multiple independent datasets. This study significantly extends the novel integrative analysis paradigm by conducting dimension reduction analysis. An important contribution is that, to promote more effectively similarity across datasets, two contrasted penalties have been developed. The sign-based contrasted penalty has not been well studied in the literature. Thus its methodological, theoretical, and numerical investigations may have independent value beyond this article. Rigorous theoretical investigation establishes the consistency property of the proposed integrative analysis. Effective computational algorithms have been developed. Extensive simulations demonstrate satisfactory performance of the proposed analysis. iSPCA$_M$ and iSPCA$_S$ do not dominate each other. Their performance depends on data/model setting. They are thus both needed in practice. In data analysis, the proposed integrative analysis leads to findings different from meta-analysis and stacked analysis. By promoting similarity across datasets, iSPCA$_M$ and iSPCA$_S$ lead to more consistent findings. The proposed integrative analysis is observed to have satisfactory stability.

Table 7
Simulation results for sSPCA. In each cell: median of the average TPR values and FDR values across the M datasets (median absolute deviation/0.6745).

β	Case 1				Case 2				Case 3			
	$d = 500$		$d = 1000$		$d = 500$		$d = 1000$		$d = 500$		$d = 1000$	
	TPR	FDR	TPR	FDR	TPR	FDR	TPR	FDR	TPR	FDR	TPR	FDR
Scenario I												
0.3	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
0.5	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
0.8	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Scenario II												
0.3	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
0.5	0.95	0.00	0.97	0.00	1.00	0.00	1.00	0.00	0.95	0.00	0.97	0.00
	(0.03)	(0.00)	(0.05)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.07)	(0.00)	(0.00)	(0.00)
0.8	0.99	0.00	0.99	0.00	0.99	0.00	0.99	0.00	0.97	0.01	0.97	0.01
	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)
Scenario III												
0.3	0.96	0.28	1.00	0.30	0.92	0.21	0.96	0.25	0.92	0.21	0.96	0.25
	(0.06)	(0.09)	(0.00)	(0.03)	(0.00)	(0.00)	(0.00)	(0.00)	(0.06)	(0.10)	(0.00)	(0.06)
0.5	0.96	0.40	0.97	0.40	0.83	0.17	0.83	0.17	0.90	0.34	0.84	0.28
	(0.03)	(0.06)	(0.04)	(0.04)	(0.00)	(0.00)	(0.00)	(0.00)	(0.04)	(0.04)	(0.07)	(0.07)
0.8	1.00	0.44	1.00	0.44	0.88	0.30	0.88	0.30	0.92	0.38	0.91	0.38
	(0.00)	(0.02)	(0.00)	(0.02)	(0.00)	(0.00)	(0.00)	(0.00)	(0.04)	(0.10)	(0.04)	(0.10)
Scenario IV												
0.3	0.17	0.75	0.18	0.75	0.25	0.75	0.25	0.75	0.21	0.75	0.21	0.75
	(0.06)	(0.00)	(0.05)	(0.00)	(0.06)	(0.00)	(0.03)	(0.00)	(0.06)	(0.00)	(0.03)	(0.00)
0.5	0.23	0.73	0.20	0.73	0.27	0.71	0.27	0.73	0.22	0.73	0.21	0.73
	(0.08)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.03)	(0.02)	(0.03)	(0.02)
0.8	0.91	0.65	0.92	0.69	0.71	0.57	0.69	0.59	0.62	0.58	0.49	0.58
	(0.05)	(0.05)	(0.05)	(0.04)	(0.03)	(0.02)	(0.01)	(0.01)	(0.17)	(0.04)	(0.14)	(0.03)

This study can be potentially extended in multiple directions. Beyond PCA, there are quite a few other dimension reduction techniques. It can be of interest to conduct integrative analysis based on other techniques. For regularized estimation and selection, we adopt the composite MCP. This penalty has demonstrated satisfactory performance in other contexts. There are other penalties that can also serve the purpose of regularized estimation and two-level selection. Both the magnitude- and sign-based contrasted penalties can have applications far beyond this study. In data analysis, the proposed integrative analysis leads to findings different from meta-analysis and stacked analysis. The satisfactory stability provides certain support to the validity of analysis. However, additional analysis is needed to confirm the findings (though there are no commonly accepted approaches for evaluating SPCA results in the literature).

Acknowledgments

We thank the Editor-in-Chief and reviewers for their careful review and insightful comments, which have led to a significant improvement of the article. This study was supported by the National Natural Science Foundation of China (71471152), National Bureau of Statistics of China (2016LD01, 2015629), and Fundamental Research Funds for the Central Universities (20720171064).

Appendix

Proof of Theorems 1 and 2

Here we prove Theorem 1. Theorem 2 then follows from Theorem 1 and similar arguments as in [17]. For each $m \in \{1, \ldots, M\}$, let $\hat{v}_1^{(m)}$ denote the first right singular vector of $X^{(m)}$. With $\tilde{v}_1^{(m)} = \hat{v}_1^{(m)}$, the penalized estimate can be rewritten as

$$\breve{\mathbf{u}}_1 = \underset{\breve{u}_1^{(m)}, m\in\{1,\ldots,M\}}{\operatorname{argmin}} \sum_{m=1}^{M} \frac{1}{2n_m} \|X^{(m)} - \breve{u}_1^{(m)} \hat{v}_1^{(m)\top}\|_F^2 + \sum_{i=1}^{d} \rho\left(\sum_{m=1}^{M} \rho(|\breve{u}_{i,1}^{(m)}|; \mu_1, a); 1, b\right)$$
$$+ \frac{\mu_2}{2} \sum_{i=1}^{d} \sum_{1\le \ell < m \le M} \left(\frac{\breve{u}_{i,1}^{(m)}}{\sqrt{\breve{u}_{i,1}^{(m)2} + \tau^2}} - \frac{\breve{u}_{i,1}^{(\ell)}}{\sqrt{\breve{u}_{i,1}^{(\ell)2} + \tau^2}} \right)^2,$$

where $\breve{\mathbf{u}}_1 = (\breve{u}_1^{(1)}, \ldots, \breve{u}_1^{(M)})$. The first term can be rewritten as

$$\sum_{m=1}^{M} \frac{1}{2n_m} \|X^{(m)} - \breve{u}_1^{(m)} \hat{v}_1^{(m)\top}\|_F^2 = \sum_{i=1}^{d} \sum_{m=1}^{M} \frac{1}{2n_m} \sum_{j=1}^{n_m} (x_{ij}^{(m)} - \breve{u}_{i,1}^{(m)} \hat{v}_{j,1}^{(m)})^2.$$

Therefore, we can optimize over each individual row of $\breve{\mathbf{u}}_1$.

By the Karush–Kuhn–Tucker (KKT) conditions, $\breve{u}_{i,1}^{(m)}$, for each $m \in \{1, \ldots, M\}$, satisfies

$$\frac{1}{n_m} \left(\breve{u}_{i,1}^{(m)} - \sum_{j=1}^{n_m} x_{i,j}^{(m)} \hat{v}_{j,1}^{(m)} \right) + \mu_{im} \operatorname{sign}(\breve{u}_{i,1}^{(m)}) + \frac{\mu_2 c_{im} \tau^2}{(\breve{u}_{i,1}^{(m)2} + \tau^2)^{3/2}} = 0 \quad \text{if } \breve{u}_{i,1}^{(m)} \neq 0,$$

where

$$\mu_{im} = \dot{\rho} \left(\sum_{\ell=1}^{M} \rho(|\breve{u}_{i,1}^{(\ell)}|; \mu_1, a), 1, b \right) \dot{\rho}(|\breve{u}_{i,1}^{(m)}|; \mu_1, a), \quad c_{im} = \sum_{\ell \neq m} \left(\breve{u}_{i,1}^{(m)} / \sqrt{\breve{u}_{i,1}^{(m)2} + \tau^2} - \breve{u}_{i,1}^{(\ell)} / \sqrt{\breve{u}_{i,1}^{(\ell)2} + \tau^2} \right).$$

In addition,

$$\Big| \sum_{j=1}^{n_m} x_{i,j}^{(m)} \hat{v}_{j,1}^{(m)} + \mu_2 \sum_{\ell \neq m} \frac{\breve{u}_{i,1}^{(\ell)}}{\tau \sqrt{\breve{u}_{i,1}^{(\ell)2} + \tau^2}} \Big| \leq \mu_1 \dot{\rho} \left(\sum_{\ell \neq m}^{M} \rho(|\breve{u}_{i,1}^{(\ell)}|; \mu_1, a), 1, b \right) \quad \text{if } \breve{u}_{i,1}^{(m)} = 0.$$

Denote $\hat{u}_{i,1}^{(m)} = \sum_{j=1}^{n_m} x_{i,j}^{(m)} \hat{v}_{j,1}^{(m)}$. Since $\mu_2 = o(\mu_1)$ and c_{im} is bounded, we have

$$\mu_2 c_{im} \tau^2 (\breve{u}_{i,1}^{(m)2} + \tau^2)^{-3/2} = o(\mu_1^{-2})$$

and

$$\dot{\rho}(|\breve{u}_{i,1}^{(m)}|; \mu_1, a) = \mu_1 (1 - |\breve{u}_{i,1}^{(m)}|/(\mu_1 a))_+ = 0$$

if $|\breve{u}_{i,1}^{(m)}| \geq \mu_1 a$. We also have $\breve{u}_{i,1}^{(m)} = \hat{u}_{i,1}^{(m)} + o(\mu_1^{-2})$ if $|\breve{u}_{i,1}^{(m)}| \geq \mu_1 a$. In addition

$$|\breve{u}_{i,1}^{(m)}| + n_m \dot{\rho} \left(\sum_{\ell=1}^{M} \rho(|\breve{u}_{i,1}^{(\ell)}|; \mu_1, a), 1, b \right) (\mu_1 - |\breve{u}_{i,1}^{(m)}|/a) = |\hat{u}_{i,1}^{(m)} + o(\mu_1)|$$

if $0 < |\breve{u}_{i,1}^{(m)}| < a\mu_1$. Without loss of generality, suppose that $n_m > a$ for all $m \in \{1, \ldots, M\}$. With $b = 1/2Ma\mu_1^2$,

$$\dot{\rho} \left(\sum_{\ell=1}^{M} \rho(|\breve{u}_{i,1}^{(\ell)}|; \mu_1, a), 1, b \right) \leq 1,$$

and

$$|\breve{u}_{i,1}^{(m)}| + n_m \dot{\rho} \left(\sum_{\ell=1}^{M} \rho(|\breve{u}_{i,1}^{(\ell)}|; \mu_1, a), 1, b \right) (\mu_1 - |\breve{u}_{i,1}^{(m)}|/a) \leq |\breve{u}_{i,1}^{(m)}| + n_m(\mu_1 - |\breve{u}_{i,1}^{(m)}|/a) \leq n_m \mu_1,$$

if $0 < |\breve{u}_{i,1}^{(m)}| < a\mu_1$. In other words, $|\breve{u}_{i,1}^{(m)}| > a\mu_1$ if $|\hat{u}_{i,1}^{(m)} + o(\mu_1)| > n_m \mu_1$. Denote $\xi_{im} = |\hat{u}_{i,1}^{(m)} + o(\mu_1)|$. Define

$$u_{i,1}^{*(m)} = \breve{u}_{i,1}^{(m)} \mathbf{1}(\xi_{im} > n_m \mu_1) + \xi_{im} \mathbf{1}(\xi_{im} \leq n_m \mu_1) \mathbf{1}\{|\hat{u}_{i,1}^{(m)}/n_m| \geq \mu_1/M - \mu_2(M-1)/\tau\},$$
$$u_{i,1}^{'(m)} = \breve{u}_{i,1}^{(m)} \mathbf{1}(\xi_{im} > n_m \mu_1) - \xi_{im} \mathbf{1}(\xi_{im} \leq n_m \mu_1) \mathbf{1}\{|\hat{u}_{i,1}^{(m)}/n_m| \geq \mu_1/M - \mu_2(M-1)/\tau\},$$

for each $i \in \{1, \ldots, d\}$. Following the proof of Theorem 3.1 in [17], we can show that $\tilde{u}_1^{*(m)} = u_1^{*(m)}/\|u_1^{*(m)}\|$ and $\tilde{u}_1^{'(m)} = u_1^{'(m)}/\|u_1^{'(m)}\|$ are consistent for $u_1^{(m)}$, with a convergence rate of $d^{\kappa/2}$.

We can see that, if $|\hat{u}_{i,1}^{(m)} + o(\mu_1)| > n_m \mu_1$, then $u_{i,1}^{'(m)} = u_{i,1}^{*(m)} = \breve{u}_{i,1}^{(m)}$. If $u_{i,1}^{'(m)} = u_{i,1}^{*(m)} = 0$, then $|\hat{u}_{i,1}^{(m)}/n_m| < \mu_1/M - \mu_2(M-1)/\tau$, which indicates that

$$|\hat{u}_{i,1}^{(m)}/n_m + \mu_2 \sum_{\ell \neq m} \left(\frac{\breve{u}_{i,1}^{(\ell)}}{\sqrt{\breve{u}_{i,1}^{(\ell)2} + \tau^2}} \right) \tau^{-1}| \leq \left| \frac{\hat{u}_{i,1}^{(m)}}{n_m} \right| + \frac{\mu_2(M-1)}{\tau} \leq \frac{\mu_1}{M}.$$

Since $\dot{\rho}(\sum_{\ell \neq m}^{M} \rho(|\breve{u}_{i,1}^{(\ell)}|; \mu_1, a), 1, b) \geq 1/M$, with $b = 1/2aM\mu_1^2$, combined with the KKT conditions, we can get that

$$|\hat{u}_{i,1}^{(m)}/n_m| < \mu_1/M - \mu_2(M-1)/\tau \quad \Rightarrow \quad \breve{u}_{i,1}^{(m)} = 0.$$

Then we can conclude that $u_{i,1}^{'(m)} \le \breve{u}_{i,1}^{(m)} \le u_{i,1}^{*(m)}$. And it is easy to see that $\tilde{u}_1^{(m)} = \breve{u}_1^{(m)}/\|\breve{u}_1^{(m)}\|$ is consistent for $u_1^{(m)}$, with a convergence rate of $d^{\kappa/2}$. □

Computational algorithm for iSPCA$_M$

Consider the penalized objective function

$$\tilde{L}(\breve{u}_1^{(1)}, \ldots, \breve{u}_1^{(M)}) = \sum_{m=1}^{M} \frac{1}{2n_m} \|X^{(m)} - \breve{u}_1^{(m)} \tilde{v}_1^{(m)\top}\|_F^2 + \sum_{i=1}^{d} \rho\left(\sum_{m=1}^{M} \rho(|\breve{u}_{i,1}^{(m)}|; \mu_1, a); 1, b\right) + \frac{\mu_2}{2} \sum_{i=1}^{d} \sum_{1\le \ell < m \le M} (\breve{u}_{i,1}^{(m)} - \breve{u}_{i,1}^{(\ell)})^2,$$
$$\text{subject to} \quad \|\tilde{v}_1^{(1)}\| = \cdots = \|\tilde{v}_1^{(M)}\| = 1.$$

We adopt a similar computational algorithm as for iSPCA$_S$. The key difference lies in Step 2(a). Consider

$$\breve{\mathbf{u}}_1 = f(\tilde{v}_1^{(1)}, \ldots, \tilde{v}_1^{(M)}) = \underset{\breve{u}_1^{(m)}, m\in\{1,\ldots,M\}}{\operatorname{argmin}} \quad \tilde{L}(\breve{u}_1^{(1)}, \ldots, \breve{u}_1^{(M)}),$$

with fixed $\tilde{v}_1^{(1)}, \ldots, \tilde{v}_1^{(M)}$. Suppose that we have initial estimate $\{\breve{u}_{i,1}^{(1)0}, \ldots, \breve{u}_{i,1}^{(M)0}\}$ which is close to the minimizer. This can be obtained from the standard SPCA. For $\breve{u}_{i,1}^{(m)}$, with fixed $\tilde{v}_1^{(1)}, \ldots, \tilde{v}_1^{(M)}$,

$$\tilde{L}(\breve{u}_{i,1}^{(m)}) \approx \frac{1}{2n_m} \sum_{j=1}^{n_m} (x_{ij}^{(m)} - \breve{u}_{i,1}^{(m)} \tilde{v}_{j,1}^{(m)})^2 + \mu_{im}(|\breve{u}_{i,1}^{(m)}| - |\breve{u}_{i,1}^{(m)0}|) + \frac{\mu_2}{2} \sum_{\ell\neq m} (\breve{u}_{i,1}^{(m)} - \breve{u}_{i,1}^{(\ell)0})^2,$$

where $\mu_{im} = \dot{\rho}(\sum_{\ell=1}^{M} \rho(|\breve{u}_{i,1}^{(\ell)0}|; \mu_1, a), 1, b)\dot{\rho}(|\breve{u}_{i,1}^{(m)0}|; \mu_1, a)$. Then

$$\breve{u}_{i,1}^{(m)} = \frac{n_m}{1 + \mu_2 n_m (M-1)} S(Z_i^{(m)0}, \mu_{im}),$$

where

$$Z_i^{(m)0} = \frac{1}{n_m} \sum_{j=1}^{n_m} x_{ij}^{(m)} \tilde{v}_{j,1}^{(m)} + \mu_2 \sum_{\ell\neq m} \breve{u}_{i,1}^{(\ell)0}.$$

Thus, in Step 2(a) we can solve for $\breve{\mathbf{u}}_{1,\text{new}}$ as follows.

For the ith row of $\breve{\mathbf{u}}_{1,\text{new}}$, with $i \in \{1, \ldots, d\}$,

1. Let $\{\breve{u}_{i,1}^{(1)0}, \ldots, \breve{u}_{i,1}^{(M)0}\} = \{\breve{u}_{i,1,\text{old}}^{(1)}, \ldots, \breve{u}_{i,1,\text{old}}^{(M)}\}$, and fix $\{\tilde{v}_1^{(1)}, \ldots, \tilde{v}_1^{(M)}\} = \{\tilde{v}_{1,\text{old}}^{(1)}, \ldots, \tilde{v}_{1,\text{old}}^{(M)}\}$.
2. For each $m \in \{1, \ldots, M\}$,
   (a) update $\mu_{im} = \dot{\rho}\left(\sum_{\ell=1}^{M} \rho(|\breve{u}_{i,1}^{(\ell)0}|; \mu_1, a), 1, b\right) \dot{\rho}(|\breve{u}_{i,1}^{(m)0}|; \mu_1, a)$;
   (b) update $\breve{u}_{i,1}^{(m)0} = n_m/(1 + \mu_2 n_m (M-1))S(Z_i^{(m)0}, \mu_{im})$.
3. Return $(\breve{u}_{i,1}^{(1)0}, \ldots, \breve{u}_{i,1}^{(M)0})$.

References

[1] J. Chiquet, Y. Grandvalet, C. Ambroise, Inferring multiple graphical structures, Stat. Comput. 21 (4) (2011) 537–553.
[2] A. d'Aspremont, L.E. Ghaoui, M.I. Jordan, G.R. Lanckriet, A direct formulation for sparse PCA using semidefinite programming, SIAM Rev. 49 (3) (2007) 434–448.
[3] R. Guerra, D.R. Goldstein, Meta-analysis and Combining Iinformation in Genetics and Genomics, CRC Press, 2009.
[4] J. Huang, S. Ma, Variable selection in the accelerated failure time model via the bridge method, Lifetime Data Anal. 16 (2) (2010) 176–195.
[5] I.M. Johnstone, A.Y. Lu, On consistency and sparsity for principal components analysis in high dimensions, J. Amer. Statist. Assoc. 104 (486) (2009) 682–693.
[6] I.T. Jolliffe, N.T. Trendafilov, M. Uddin, A modified principal component technique based on the lasso, J. Comput. Graph. Statist. 12 (3) (2003) 531–547.
[7] S. Jung, J.S. Marron, PCA consistency in high dimension, low sample size context, Ann. Statist. 37 (6B) (2009) 4104–4130.
[8] M.R. Leadbetter, G. Lindgren, H. Rootzén, Extremes and Related Properties of Random Sequences and Processes, Springer-Verlag, New York, 1983.
[9] J. Liu, J. Huang, S. Ma, Integrative analysis of multiple cancer genomic datasets under the heterogeneity model, Stat. Med. 32 (20) (2013) 3509–3521.
[10] J. Liu, J. Huang, Y. Zhang, Q. Lan, N. Rothman, T. Zheng, S. Ma, Integrative analysis of prognosis data on multiple cancer subtypes, Biometrics 70 (3) (2014) 480–488.
[11] J. Liu, S. Ma, J. Huang, Integrative analysis of cancer diagnosis studies with composite penalization, Scand. J. Stat. 41 (1) (2014) 87–103.
[12] Z. Ma, Sparse principal component analysis and iterative thresholding, Ann. Statist. 41 (2) (2013) 772–801.
[13] S. Ma, Y. Dai, Principal component analysis based methods in bioinformatics studies, Brief. Bioinform. 12 (6) (2011) 714–722.
[14] S. Ma, J. Huang, M.S. Moran, Identification of genes associated with multiple cancers via integrative analysis, BMC Genomics 10 (1) (2009) 535.
[15] D. Paul, Asymptotics of sample eigenstructure for a large dimensional spiked covariance model, Statist. Sinica 17 (4) (2007) 1617–1642.
[16] H. Shen, J.Z. Huang, Sparse principal component analysis via regularized low rank matrix approximation, J. Multivariate Anal. 99 (6) (2008) 1015–1034.

[17] D. Shen, H. Shen, J.S. Marron, Consistency of sparse PCA in high dimension, low sample size contexts, J. Multivariate Anal. 115 (2013) 317–333.
[18] D. Shen, H. Shen, H. Zhu, J.S. Marron, The statistics and mathematics of high dimension low sample size asymptotics, Statist. Sinica 26 (4) (2016) 1747–1770.
[19] X. Shi, J. Liu, J. Huang, Y. Zhou, B. Shia, S. Ma, Integrative analysis of high-throughput cancer studies with contrasted penalization, Genet. Epidemiol. 38 (2) (2014) 144–151.
[20] X. Shi, S. Shen, J. Liu, J. Huang, Y. Zhou, S. Ma, Similarity of markers identified from cancer gene expression studies: observations from GEO, Brief. Bioinform. 15 (5) (2014) 671–684.
[21] L. Tang, P. Song, Fused lasso approach in regression coefficients clustering: Learning parameter heterogeneity in data integration, J. Mach. Learn. Res. 17 (2016) 3815–3937.
[22] G.C. Tseng, D. Ghosh, X.J. Zhou, Integrating Omics Data, Cambridge University Press, 2015.
[23] F. Wang, L. Wang, P. Song, Fused lasso with the adaptation of parameter ordering in combining multiple studies with repeated measurements, Biometrics 72 (2016) 184–193.
[24] D.M. Witten, R.J. Tibshirani, T. Hastie, A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, Biostatistics 10 (3) (2009) 515–534.
[25] C.-H. Zhang, Nearly unbiased variable selection under minimax concave penalty, Ann. Statist. 38 (2) (2010) 894–942.
[26] Q. Zhao, X. Shi, J. Huang, J. Liu, Y. Li, S. Ma, Integrative analysis of -omics data using penalty functions, Wiley Interdiscip. Rev. Comput. Stat. 7 (1) (2015) 99–108.
[27] H. Zou, T. Hastie, R.J. Tibshirani, Sparse principal component analysis, J. Comput. Graph. Statist. 15 (2) (2006) 265–286.