JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2020, VOL. 29, NO. 3, 592–607
https://doi.org/10.1080/10618600.2020.1713798

Taylor & Francis
Taylor & Francis Group

Diagonally Dominant Principal Component Analysis

Zheng Tracy Ke[a], Lingzhou Xue[b], and Fan Yang[c]

[a]Department of Statistics, Harvard University, Cambridge, MA; [b]Department of Statistics, Pennsylvania State University, State College, PA; [c]Department of Statistics, University of Chicago, Chicago, IL

ABSTRACT

We consider the problem of decomposing a large covariance matrix into the sum of a low-rank matrix and a diagonally dominant matrix, and we call this problem the "diagonally dominant principal component analysis (DD-PCA)." DD-PCA is an effective tool for designing statistical methods for strongly correlated data. We showcase the use of DD-PCA in two statistical problems: covariance matrix estimation and global detection in multiple testing. Using the output of DD-PCA, we propose a new estimator for estimating a large covariance matrix with factor structure. Thanks to a nice property of diagonally dominant matrices, this estimator enjoys the advantage of simultaneous good estimation of the covariance matrix and the precision matrix (by a plain inversion). A plug-in of this estimator to linear discriminant analysis and portfolio optimization yields appealing performance in real data. We also propose two new tests for testing the global null hypothesis in multiple testing when the *z*-scores have a factor covariance structure. Both tests first use DD-PCA to adjust the individual *p*-values and then plug in the adjusted *p*-values to the higher criticism (HC) test. These new tests significantly improve over the HC test and compare favorably with other existing tests. For computation of DD-PCA, we propose an iterative projection algorithm and an ADMM algorithm. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received May 2019
Revised November 2019

KEYWORDS
Alternating projection; Approximate factor model; Covariance matrix estimation; Decorrelation; Higher criticism; POET

1. Introduction

The *approximate low-rankness* is a popular structural assumption on covariance matrices. It assumes that a $p \times p$ covariance matrix $\boldsymbol{\Sigma}$ decomposes into

$$\boldsymbol{\Sigma} = \boldsymbol{L} + \boldsymbol{A}, \qquad \text{where} \quad \text{rank}(\boldsymbol{L}) = K \ll p, \\ \text{and } \boldsymbol{A} \text{ is a "nice" matrix.} \tag{1}$$

Equivalently, it introduces a latent factor model on any random vector X whose covariance matrix is $\boldsymbol{\Sigma}$, where $\boldsymbol{A}$ is the "residual covariance matrix" after the effects of latent variables are removed. Such a decomposition is not unique and varies with the meaning of a "nice" $\boldsymbol{A}$. One can impose different requirements on $\boldsymbol{A}$ to facilitate different applications. In the classical factor models for econometrics and finance, $\boldsymbol{A}$ is assumed a diagonal matrix (Fama and French 1993) or a sparse matrix (Chamberlain and Rothschild 1983), to enforce that the idiosyncratic noise accounts for little cross-sectional risk. In large-scale multiple testing, it is often assumed that the covariance matrix of test statistics has the above decomposition with $\boldsymbol{A}$ being a diagonal matrix (Leek and Storey 2008) or having a small Frobenius norm (Fan, Han, and Gu 2012). The motivation there is development of factor-adjusted multiple testing procedures, to make it legitimate to use conventional multiple testing methods on the post-factor-removal data. In image processing, a similar decomposition on image matrices was proposed (Candès et al. 2011), where $\boldsymbol{A}$ is assumed sparse, for the purpose of capturing details of images. In this article, we explore a new type of *approximate low-rankness* where

$$\text{Each diagonal of } \boldsymbol{A} \text{ is large compared with} \\ \text{other entries in the same row.} \tag{2}$$

Translated to the latent variable representation, it means, after the effects of latent variables are removed, the *correlation* matrix of "residual" variables have uniformly small off-diagonal entries. One motivation of imposing this condition is to take into account the varying scale of the diagonal elements of $\boldsymbol{A}$. Most aforementioned approximate low-rank decompositions first perform PCA on $\boldsymbol{\Sigma}$ (or an empirical version of it) to remove the first a few principal components, and then conduct operations on the remaining matrix. It is often observed that the diagonal elements of the remaining matrix have considerable variations in magnitude. To deal with it requires careful adjustment on the post-PCA operations, such as adaptive thresholding (Cai and Liu 2011). On the contrary, we impose the requirement (2) directly in the decomposition (1), in hopes of improving the PCA factors and easing the post-PCA operations. Another motivation of adopting the assumption (2) is to guarantee that $\boldsymbol{A}^{-1}$ is well-behaved. In many applications such as portfolio management and linear discriminant analysis, $\boldsymbol{A}^{-1}$ plays a key role (see Section 2). In the decomposition (1), forcing $\boldsymbol{A}$ to be a strictly diagonal matrix can ensure both $\boldsymbol{A}$ and $\boldsymbol{A}^{-1}$ are well-behaved, but this requirement is often too restrictive, and (2) is a natural relaxation. We note that imposing the

CONTACT Zheng Tracy Ke zke@fas.harvard.edu Department of Statistics, Harvard University, 1 Oxford Street, Science Center Suite 400, Cambridge, MA 02138.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JCGS.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

common sparsity assumption on $\boldsymbol{A}$ does not even guarantee positive definiteness. Despite of remedies such as increasing the threshold or projection to the positive semidefinite cone (Fan, Liao, and Liu 2016), these approaches still do not guarantee that $\boldsymbol{A}^{-1}$ is a "nice" matrix.

To formulate (2) mathematically, we define the set of "symmetric c-diagonally dominant" matrices, for any $c > 0$:

$$\mathcal{SDD}_c^+ \\ = \left\{\boldsymbol{A} = (a_{ij})_{p\times p} : \boldsymbol{A}^T = \boldsymbol{A},\ a_{jj} \geq c \sum_{i:i\neq j} |a_{ji}| \text{ for all } 1 \leq j \leq p\right\}. \tag{3}$$

For $c = 1$, it reduces to the usual definition of diagonally dominant matrices, and we omit the subscript and write $\mathcal{SDD}_1^+ = \mathcal{SDD}^+$. Given a $p \times p$ positive semidefinite matrix $\boldsymbol{S}$, we introduce an optimization problem:

$$\min_{(\boldsymbol{L},\boldsymbol{A})} \|\boldsymbol{S} - \boldsymbol{L} - \boldsymbol{A}\|_F, \qquad \text{subject to} \quad \text{rank}(\boldsymbol{L}) \leq K, \\ \boldsymbol{L} = \boldsymbol{L}^T, \quad \boldsymbol{A} \in \mathcal{SDD}_c^+, \tag{4}$$

where $\|\cdot\|_F$ is the matrix Frobenius norm. We call it *diagonally dominant principal component analysis* (*DD-PCA*). In this article, we are primarily interested in $c = 1$; discussions of $c \neq 1$ are deferred to Section 4.

The definition of DD-PCA is a nonconvex optimization with a rank constraint. Similar to solving other rank constrained optimizations in matrix completion, one can either solve a convex relaxation of (4) or develop an iterative algorithm that converges to a local minimum of (4). These ideas generate several variants of DD-PCA, as detailed in Section 4. Among those variants, one is of particular interest, which we call *one-step DD-PCA*:

- PCA: Obtain the K leading eigenvalues and eigenvectors of $\boldsymbol{S}$, denoted as $\lambda_1 \geq \cdots \geq \lambda_K \geq 0$ and $\xi_1, \ldots, \xi_K \in \mathbb{R}^p$. Let $\boldsymbol{L} = \sum_{k=1}^K \lambda_k \xi_k \xi_k^T$.
- Projection to $\mathcal{SDD}^+$: Initialize $\boldsymbol{A}^{(0)} = \boldsymbol{S} - \boldsymbol{L}$ and $\boldsymbol{J}^{(0)} = \boldsymbol{0}$. For $t = 1, 2, \ldots,$
  - Run the MRT algorithm (Mendoza, Raydan, and Tarazaga 1998)[1] to project $[\boldsymbol{A}^{(t-1)} - \boldsymbol{J}^{(t-1)}]$ into the diagonally dominant cone. Let $\boldsymbol{G}^{(t)}$ be the projected matrix.
  - Update $\boldsymbol{A}^{(t)} = \frac{1}{2}[\boldsymbol{G}^{(t-1)} + (\boldsymbol{G}^{(t-1)})^T]$ and $\boldsymbol{J}^{(t)} = \boldsymbol{J}^{(t-1)} + (\boldsymbol{G}^{(t)} - \boldsymbol{A}^{(t-1)})$.
  - If $\|\boldsymbol{J}^{(t)} - \boldsymbol{J}^{(t-1)}\|_F \leq \epsilon$, stop and output $\boldsymbol{A} = \boldsymbol{A}^{(t)}$.

This method is obtained by running one outer-loop iteration in the iterative algorithm to be introduced in Section 4, explaining the name of *one-step DD-PCA*. It has the same philosophy as the one-step Huber estimator (Bickel 1975) and one-step LLA implementation of nonconvex penalized linear regressions (Zou and Li 2008; Fan, Xue, and Zou 2014). It provides an approximate solution to (4), which is much faster to compute than solving (4) exactly.

Exploring the approximate low-rank structures is a powerful strategy for big data analysis. The classical PCA has motivated many statistical methods. Similarly, DD-PCA and one-step DD-PCA can also serve as building blocks for statistical methodology development. We exemplify it in two statistical problems: the first is estimating a large covariance matrix, and the second is testing of the global null hypothesis in multiple testing.

Estimation of large covariance matrices is a popular topic in statistical literatures (Fan, Liao, and Liu 2016). At the heart of it is two fundamental questions: (a) What structural assumption is appropriate? (b) How to evaluate the methods in real applications?

We adopt the structural assumption that the true covariance matrix $\boldsymbol{\Sigma}$ has an approximate low-rank decomposition with $\boldsymbol{A} \in \mathcal{SDD}_c^+$. This is a special type of factor covariance structures that are commonly used in econometrics (Fan and Fan 2008), finance (Fama and French 1993), genetics (Price et al. 2006), and many other fields. Our work is unique in the diagonal dominance assumption on $\boldsymbol{A}$. Intuitively, it is a natural relaxation of assuming $\boldsymbol{A}$ is diagonal, and it implies that, after the effects of latent factors are removed, the "residual variables" are almost *uncorrelated*. Compared with existing covariance matrix estimators that assume $\boldsymbol{A}$ is sparse (e.g., Fan, Liao, and Mincheva 2011, 2013), this diagonal dominance structure benefits simultaneous estimation of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$: In factor covariance structures, the singular values of the low-rank matrix are much larger than $\|\boldsymbol{A}\|$, so the error of estimating $\boldsymbol{\Sigma}$ is dominated by the error of recovering the low-rank part. If our goal is merely to estimate $\boldsymbol{\Sigma}$, we do not gain much from exploring the diagonal dominance structure of $\boldsymbol{A}$. However, if we are also interested in estimating $\boldsymbol{\Sigma}^{-1}$, the error of estimating $\boldsymbol{A}^{-1}$ will play a key role. Note that there always exists a matrix $\boldsymbol{B} \in \mathbb{R}^{n\times K}$ such that $\boldsymbol{L} = \boldsymbol{B}\boldsymbol{B}^T$. It follows from the matrix inverse formula (Horn and Johnson 2012) that

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{I}_K + \boldsymbol{B}^T\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{A}^{-1}.$$

Suppose we have obtained a good estimator $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^T + \widehat{\boldsymbol{A}}$ by fitting some factor covariance structure on the data. Even though $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|$ is small, it is still possible that $\|\widehat{\boldsymbol{A}}^{-1} - \boldsymbol{A}^{-1}\|$ is large so that $\widehat{\boldsymbol{\Sigma}}^{-1}$ is far from being a good estimator of $\boldsymbol{\Sigma}^{-1}$. Fortunately, exploring the diagonal dominance structure largely mitigates this issue, thanks to an appealing feature of the diagonally dominant cone $\mathcal{SDD}_c^+$ (Horn and Johnson 2012):

$$\|\boldsymbol{A}^{-1}\| \leq \frac{c}{c-1}\|[\text{diag}(\boldsymbol{A})]^{-1}\|, \\ \text{for any } \boldsymbol{A} \in \mathcal{SDD}_c^+, \text{ where } c > 1.$$

Therefore, if we enforce $\widehat{\boldsymbol{A}} \in \mathcal{SDD}_c^+$ in fitting the factor covariance structure, for a constant $c > 1$, then $\|\widehat{\boldsymbol{A}}^{-1}\|$ will not explode, preventing ill behavior of $\widehat{\boldsymbol{\Sigma}}^{-1}$. To this end, we propose a new covariance matrix estimator $\widehat{\boldsymbol{\Sigma}}_{\text{ddpca}}$ using the solution of DD-PCA or one-step DD-PCA. We demonstrate in numerical studies: $\widehat{\boldsymbol{\Sigma}}_{\text{ddpca}}$ has comparable performance with state-of-art methods, but the new estimator is tuning free once K is specified, so is more convenient to use. Moreover, when it comes to estimating $\boldsymbol{\Sigma}^{-1}$ by $\widehat{\boldsymbol{\Sigma}}_{\text{ddpca}}^{-1}$, the new estimator is significantly better than inverting other factor-based covariance matrix estimators.

In real applications, estimating the covariance matrix is rarely the ultimate goal. Often, it serves as an intermediate step for

[1] The MRT algorithm computes the unique projection of a $p \times p$ matrix to the convex polyhedral cone consisting of all diagonally dominant matrices. It has a complexity of $O(p^2 \log(p))$. See Section 4.

downstream tasks. We demonstrate the usefulness of our covariance estimator by evaluating its performance in two downstream tasks, portfolio management and linear discriminant analysis. In the former, an estimate of the covariance matrix is needed to obtain Markowitz's optimal portfolio weights; in the latter, it is used to compute Fisher's LDA classifier. Note that what is actually plugged into these downstream tasks is the *inverse* of estimated covariance matrix. As we have argued, the main advantage of our method is on estimating $\mathbf{\Sigma}^{-1}$ by $\widehat{\mathbf{\Sigma}}^{-1}_{\text{ddpca}}$, a perfect match to these applications. This is supported by encouraging real data results. It is worthwhile mentioning that our approach is different from the approach of plugging in an existing precision matrix estimator (e.g., the graphical lasso; Friedman, Hastie, and Tibshirani 2008). These methods assume $\mathbf{\Sigma}^{-1}$ is sparse, while we assume a factor structure on $\mathbf{\Sigma}$. For portfolio data, adopting a factor covariance structure is the common practice. For classification problems, there are also many real datasets for which it is appropriate to assume a factor structure (see Section 2).

The second application of DD-PCA is for testing of the global null hypothesis in multiple testing. We are primarily interested in the setting where the mean vector in the alternative hypothesis is sparse. The higher criticism (HC) test (Donoho and Jin 2004) is known to enjoy theoretical optimality and has gained increasing popularity in applications (Wu et al. 2011; Donoho and Jin 2015). However, the orthodox HC test assumes that the individual z-scores are mutually independent. There is limited understanding of how to extend the HC test to the case where z-scores share common latent factors. We model that the covariance matrix of z-scores, $\mathbf{\Sigma}$, has a low rank plus diagonal dominance decomposition. We propose two variants of HC for this setting, both utilizing DD-PCA as a module. The first test is a modification of the innovated HC test (Hall and Jin 2010) by plugging in the estimator of $\mathbf{\Sigma}^{-1}$ from DD-PCA. The second test uses the low-rank matrix $\widehat{\boldsymbol{L}}$ from DD-PCA to remove the effects of latent factors and then applies the orthodox HC test. Both tests significantly improve over the orthodox HC test and the innovated HC test in simulations. The rationale of these testing ideas is to "transform" and "decorrelate" the marginal z-scores. The first test extends the innovated transformation of z-scores (Hall and Jin 2010), $X \mapsto \mathbf{\Sigma}^{-1}X$, from the case where $\mathbf{\Sigma}^{-1}$ is sparse to the case where $\mathbf{\Sigma}$ has a factor structure. The second test aims to estimate and remove the latent factors from z-scores so that the "residuals" are almost uncorrelated (i.e., their covariance matrix is close to being diagonal). A similar idea was briefly mentioned in Fan, Han, and Gu (2012) (called *factor-adjusted z-scores*), but it has never been used in the global testing problem. In both new tests, we can replace DD-PCA by the classical PCA, but the numerical performance will deteriorate. This indicates that exploring the diagonal dominance structure is beneficial.

The remaining of this article is organized as follows. In Section 2, we introduce a new covariance matrix estimator powered by DD-PCA, and discuss its applications in portfolio optimization (Section 2.1) and linear discriminant analysis (Section 2.2). In Section 3, we propose two new test statistics, using DD-PCA as a building block, for testing of the global null hypothesis in multiple testing. In Section 4, we address the computation of DD-PCA, by introducing an ADMM algorithm and an iterative projection algorithm for conducting the decomposition (1) and (2) for any given covariance matrix. Section 5 contains simulations and Section 6 contains concluding remarks.

2. Estimating Large Covariance Matrices by DD-PCA

Let $X \in \mathbb{R}^p$ be a multivariate random vector with a covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$, where p is presumably much larger than n. We adopt a factor model:

$$X(j) = \sum_{k=1}^{K} b_k(j) W_k + Z(j), \qquad 1 \le j \le p, \tag{5}$$

where $W_1, \dots, W_K$ are unobserved random variables (factors), $b_k \in \mathbb{R}^p$ is a nonrandom vector containing the loadings of the kth factor, and $Z \in \mathbb{R}^p$ is a random vector independent of the factors such that

$$\boldsymbol{A} \equiv \text{cov}(Z) \in \mathcal{SDD}^+. \tag{6}$$

Given iid data $X_1, \dots, X_n \in \mathbb{R}^p$, we are interested in estimating $\mathbf{\Sigma}$ and $\mathbf{\Sigma}^{-1}$.

By Models (5) and (6), the covariance matrix of X has a decomposition

$$\mathbf{\Sigma} = \boldsymbol{B}\text{cov}(W)\boldsymbol{B}^T + \boldsymbol{A}, \qquad \text{where} \quad \text{rank}\big(\boldsymbol{B}\text{cov}(W)\boldsymbol{B}^T\big) = K \quad \text{and} \quad \boldsymbol{A} \in \mathcal{SDD}^+.$$

It has the low-rank plus diagonal dominance structure. We propose the following estimator: Let $\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T$ be the sample covariance matrix. Take $\boldsymbol{S}$ as the input to the one-step DD-PCA algorithm in Section 1 and let $(\widehat{\boldsymbol{L}}, \widehat{\boldsymbol{A}})$ be the output. We estimate $\mathbf{\Sigma}$ by

$$\widehat{\mathbf{\Sigma}}_{\text{ddpca}} = \widehat{\boldsymbol{L}} + \widehat{\boldsymbol{A}}, \qquad \text{where } (\widehat{\boldsymbol{L}}, \widehat{\boldsymbol{A}}) \text{ is the output of one-step DD-PCA.} \tag{7}$$

We then estimate $\mathbf{\Sigma}^{-1}$ by the inverse of $\widehat{\mathbf{\Sigma}}_{\text{ddpca}}$. Here, $(\widehat{\boldsymbol{L}}, \widehat{\boldsymbol{A}})$ can be replaced by the output of other variants of DD-PCA (see Section 4). They give similar numerical performance, so we stick to one-step DD-PCA for computational convenience.

Different from existing covariance estimation methods under factor structures, our approach imposes the diagonal dominance constraint on $\boldsymbol{A}$. We now compare it with methods that impose the sparsity constraint on $\boldsymbol{A}$. One popular method is POET (Fan, Liao, and Mincheva 2013). Let $\boldsymbol{S} = \sum_{k=1}^{p} \lambda_k \xi_k \xi_k^T$ be the eigen-decomposition of $\boldsymbol{S}$, where λ_k and ξ_k are the kth eigenvalue and eigenvector, respectively. The POET estimator is

$$\widehat{\mathbf{\Sigma}}_{\text{poet}} = \widehat{\boldsymbol{L}}_* + \mathcal{T}_a(\widehat{\boldsymbol{A}}_*), \qquad \text{where} \quad \widehat{\boldsymbol{L}}_* = \sum_{k=1}^{K} \lambda_k \xi_k \xi_k^T, \quad \widehat{\boldsymbol{A}}_* = \sum_{k=K+1}^{p} \lambda_k \xi_k \xi_k^T. \tag{8}$$

Here, $\mathcal{T}_a(\cdot)$ can be any entry-wise adaptive thresholding operator (Rothman, Levina, and Zhu 2009; Cai and Liu 2011; Xue, Ma, and Zou 2012). Fan, Liao, and Mincheva (2013) suggested using the hard-thresholding operator applied to a "correlation matrix" associated with $\widehat{\boldsymbol{A}}_*$, that is,

$$\mathcal{T}_a(\widehat{\boldsymbol{A}}_*) = \widehat{\boldsymbol{D}}^{\frac{1}{2}} H_a\left(\widehat{\boldsymbol{D}}^{-\frac{1}{2}}\widehat{\boldsymbol{A}}_*\widehat{\boldsymbol{D}}^{-\frac{1}{2}}\right)\widehat{\boldsymbol{D}}^{\frac{1}{2}} \qquad \text{where} \quad \widehat{\boldsymbol{D}} = \mathrm{diag}(\widehat{\boldsymbol{A}}_*), \tag{9}$$

and $H_a(\cdot)$ is the entry-wise hard-thresholding at the threshold $a > 0$. Then, an estimate of $\boldsymbol{\Sigma}^{-1}$ is obtained by $\widehat{\boldsymbol{\Sigma}}_{\text{poet}}^{-1}$.

Figure 1 gives a simulation example. Fix $(p, n, K) = (2000, 200, 3)$. We generate data from the model (5), where the factors $\{W_k(i) : 1 \leq k \leq K, 1 \leq i \leq n\}$ are drawn iid from $N(0, 1)$, the factor loadings $\{b_k(j) : 1 \leq k \leq K, 1 \leq j \leq p\}$ are generated iid from $\mathcal{N}(0, 1)$, and the noise vectors $Z_1, \ldots, Z_n$ are drawn iid from a multivariate normal $\mathcal{N}_p(\mathbf{0}, \boldsymbol{A})$, with $A(i,j) = 0.5^{|i-j|+1}$ for $i \neq j$ and 1 otherwise. For both methods, K is unknown and treated as a tuning integer. POET has an additional tuning threshold a, which is selected by cross-validation (default procedure in the *poet* package).[2]

In the top four panels of Figure 1, we show the average estimation errors on $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$ over 100 repetitions. Since K is unknown, we implement both methods for the true $K = 3$ and misspecified $K \in \{4, 5, \ldots, 8\}$.[3] For estimating $\boldsymbol{\Sigma}$, the two methods give very similar performance. This is not surprising. Since the eigenvalues of the low-rank part are much larger than $\|\boldsymbol{A}\|$, the error of estimating $\boldsymbol{\Sigma}$ is dominated by the error of recovering the low-rank part. Our method and POET has the same low-rank part (the $\widehat{\boldsymbol{L}}$ from one-step DD-PCA and the $\widehat{\boldsymbol{L}}_*$ in (8) are indeed the same), so they have similar errors on estimating $\boldsymbol{\Sigma}$. From the bottom left two panels of Figure 1, we

[2] This default procedure guarantees that $\widehat{\boldsymbol{\Sigma}}_{\text{poet}}$ is invertible.

[3] We do not include the results of $K \in \{1, 2\}$, as the errors are much larger.

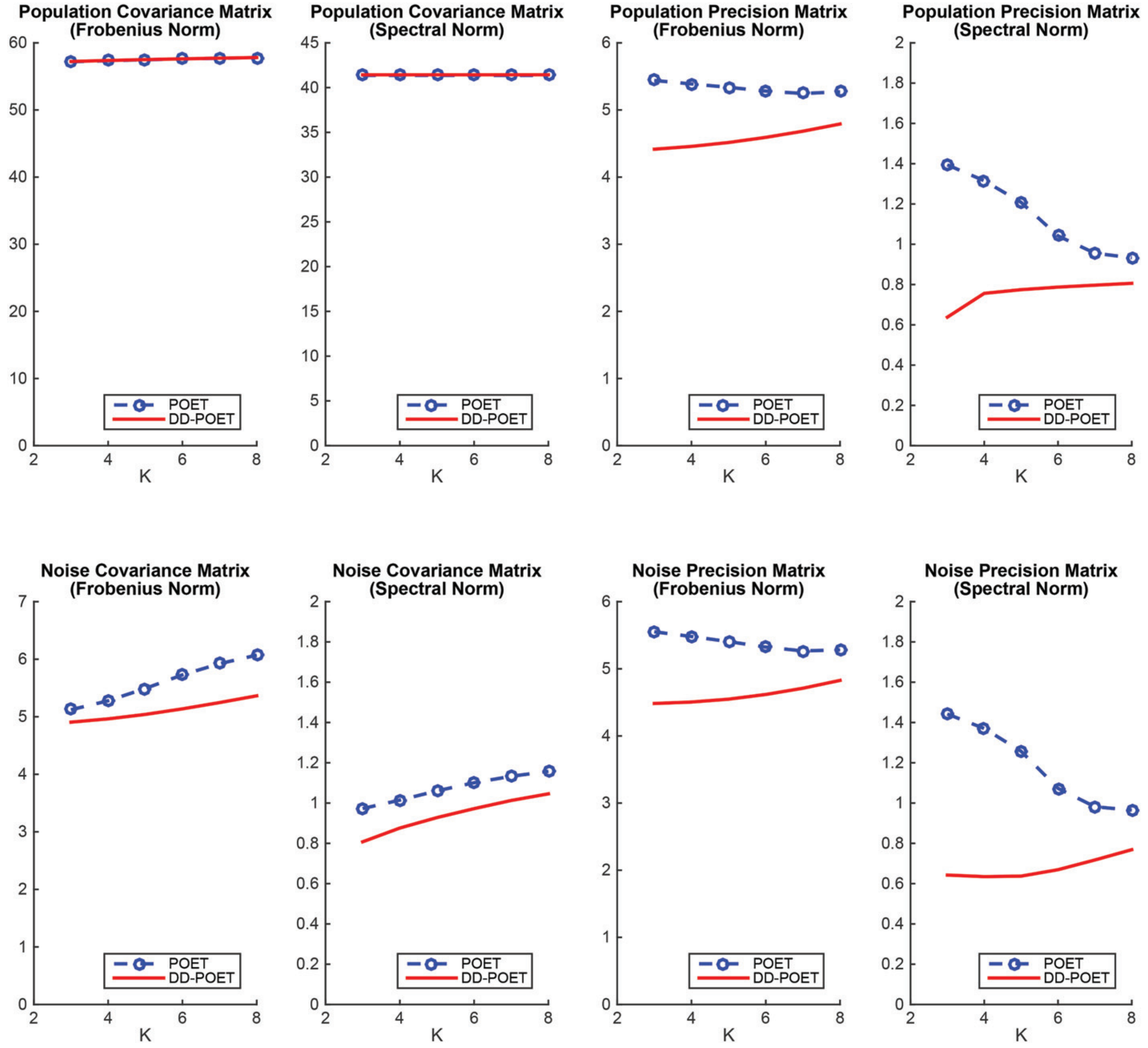

Figure 1. Comparison of our method with POET on estimating $\boldsymbol{\Sigma}$ (covariance matrix), $\boldsymbol{\Sigma}^{-1}$ (precision matrix), $\boldsymbol{A}$ (noise covariance matrix), and $\boldsymbol{A}^{-1}$ (noise precision matrix).

can see that our method does a better job on estimating $\boldsymbol{A}$; especially, the spectral norm error is 10–20% smaller. However, this improvement is almost negligible compared with the errors on recovering the low-rank part. We conclude that our method and POET have similar performance on estimating $\boldsymbol{\Sigma}$. Still, our method has an advantage: It has no tuning threshold and is more convenient to use.

How about the performance on estimating $\boldsymbol{\Sigma}^{-1}$? The top right two panels of Figure 1 clearly suggest that our method has a significant advantage. When $K = 3$, the spectral norm error of our method is only one half of the error of POET. Interestingly, the performance of POET improves with an overshooting K; but even for $K = 8$, its spectral norm error is still 10% larger than the error of our method. For the Frobenius norm error, our estimator also outperforms POET for all choices of K. This phenomenon is due to that $\widehat{\boldsymbol{A}}$ plays a dominating role when we compute the inverse of $\widehat{\boldsymbol{\Sigma}}$, while the low-rank part has a negligible effect, so the advantage of our method on recovering $\boldsymbol{A}$ becomes prominent. This is illustrated in the bottom right two panels of Figure 1. Recall that $\widehat{\boldsymbol{A}}$ is from one-step DD-PCA and $\widehat{\boldsymbol{A}}_*$ is as in (8). The Frobenius/spectral norm of $(\widehat{\boldsymbol{A}}^{-1} - \boldsymbol{A}^{-1})$ is significantly smaller than the Frobenius/spectral norm of $(\widehat{\boldsymbol{A}}_*^{-1} - \boldsymbol{A}^{-1})$. Additionally, by comparing the top right two panels with the bottom right two panels, we can see that the error of estimating $\boldsymbol{\Sigma}^{-1}$ is almost determined by the error of estimating $\boldsymbol{A}^{-1}$.

This numerical example delivers two messages: First, compared with competitive factor-based methods, the major advantage of our method is on estimating $\boldsymbol{\Sigma}^{-1}$ by $\widehat{\boldsymbol{\Sigma}}^{-1}$. Second, such an advantage is driven by the better accuracy on recovering $\boldsymbol{A}^{-1}$. Below, we explain them using linear algebra.

Without loss of generality, in model (5), we assume the covariance matrix of W equals to the identify matrix. Then, $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}^T + \boldsymbol{A}$. By matrix inverse formula,

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{I}_K + \boldsymbol{B}^T\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{A}^{-1}.$$

Suppose we construct an estimator $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^T + \widehat{\boldsymbol{A}}$ from fitting a factor-type covariance structure. Then, $\widehat{\boldsymbol{\Sigma}}^{-1}$ (if it exists) has a similar decomposition:

$$\widehat{\boldsymbol{\Sigma}}^{-1} = \widehat{\boldsymbol{A}}^{-1} - \widehat{\boldsymbol{A}}^{-1}\widehat{\boldsymbol{B}}(\boldsymbol{I}_K + \widehat{\boldsymbol{B}}^T\widehat{\boldsymbol{A}}^{-1}\widehat{\boldsymbol{B}})^{-1}\widehat{\boldsymbol{B}}^T\widehat{\boldsymbol{A}}^{-1}.$$

By some basic linear algebra, we can derive the following proposition:

Proposition 1. Let $\widehat{\boldsymbol{A}}^{-\frac{1}{2}}\widehat{\boldsymbol{B}} = \sum_{k=1}^K \hat{\sigma}_k \hat{\eta}_k \hat{h}_k'$ be the singular value decomposition of $\widehat{\boldsymbol{A}}^{-\frac{1}{2}}\widehat{\boldsymbol{B}}$, where $\hat{\sigma}_k > 0$ is the kth singular value and $\hat{\eta}_k \in \mathbb{R}^p$ and $\hat{h}_k \in \mathbb{R}^K$ are the corresponding left and right singular vectors. Then,

$$\widehat{\boldsymbol{\Sigma}}^{-1} = \widehat{\boldsymbol{A}}^{-1} - \widehat{\boldsymbol{A}}^{-\frac{1}{2}}\left(\sum_{k=1}^K \frac{1}{\hat{\sigma}_k^{-2} + 1}\hat{\eta}_k\hat{\eta}_k'\right)\widehat{\boldsymbol{A}}^{-\frac{1}{2}}. \tag{10}$$

By (10), the error of recovering the low-rank part only affects the matrix in the brackets. For $(\widehat{\boldsymbol{A}}, \widehat{\boldsymbol{B}})$ obtained in factor-based methods, nonzero eigenvalues of $\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^T$ are much larger than $\|\widehat{\boldsymbol{A}}\|$, so $\widehat{\sigma}_k$'s are all very large. Then, the matrix in the brackets can hardly bring in a large error in $\widehat{\boldsymbol{\Sigma}}^{-1}$. The error in $\widehat{\boldsymbol{\Sigma}}^{-1}$ mainly comes from the error in $\widehat{\boldsymbol{A}}^{-1}$.

We further investigate the error in $\widehat{\boldsymbol{A}}^{-1}$. Note that

$$\|\widehat{\boldsymbol{A}}^{-1} - \boldsymbol{A}^{-1}\| \le \|\widehat{\boldsymbol{A}}^{-1}\|\|\boldsymbol{A}^{-1}\|\|\widehat{\boldsymbol{A}} - \boldsymbol{A}\|. \tag{11}$$

To achieve a small $\|\widehat{\boldsymbol{A}} - \boldsymbol{A}\|$ by imposing structural assumptions on $\boldsymbol{A}$ is not too difficult. However, it typically does not prevent $\|\widehat{\boldsymbol{A}}^{-1}\|$ from exploding. For example, if $\widehat{\boldsymbol{A}}$ is obtained from entry-wise thresholding, we need a comparably large threshold to control $\|\widehat{\boldsymbol{A}}^{-1}\|$, but unfortunately we cannot let the threshold be too large as it significantly increases $\|\widehat{\boldsymbol{A}} - \boldsymbol{A}\|$. It turns out that, if we restrict $\widehat{\boldsymbol{A}} \in \mathcal{SDD}_c^+$ for a constant $c > 1$, then it is automatically guaranteed that $\|\widehat{\boldsymbol{A}}^{-1}\|$ has a nice bound. As a property of diagonally dominant matrices (Horn and Johnson 2012), for $c > 1$,

$$\begin{aligned}\lambda_{\min}(\widehat{\boldsymbol{A}}) &\ge \min_{1\le j\le p}\Big\{\widehat{a}_{jj} - \sum_{i:i\neq j}|\widehat{a}_{ji}|\Big\}\\ &\ge \min_{1\le j\le p}\Big\{\widehat{a}_{jj} - c^{-1}\widehat{a}_{jj}\Big\} \ge \frac{c-1}{c}\min_{1\le j\le p}\widehat{a}_{jj}.\end{aligned}$$

It follows that

$$\|\widehat{\boldsymbol{A}}^{-1}\| \le \frac{c}{c-1}\|[\mathrm{diag}(\widehat{\boldsymbol{A}})]^{-1}\|. \tag{12}$$

This explains why the constraint of $\widehat{\boldsymbol{A}} \in \mathcal{SDD}_c^+$ helps significantly reduce the errors in $\widehat{\boldsymbol{A}}^{-1}$ and (ultimately) the errors in $\widehat{\boldsymbol{\Sigma}}^{-1}$.

The above argument applies to $c > 1$. In our method, $c = 1$. Sometimes, we may even have to use $c < 1$, so that the assumption $\boldsymbol{A} \in \mathcal{SDD}_c^+$ is not too restrictive (see Section 4). For $c \le 1$, we do not have a solid argument as (12), but a similar phenomenon is observed in numerical studies.

Below, we use two real applications to further demonstrate that exploring the diagonal dominance factor structures is a useful strategy.

2.1. Application to Portfolio Management

Given a collection of p assets, portfolio management aims to determine the weights allocated to each asset. It is often desirable to construct the *minimum risk portfolio*, where the asset weights $\boldsymbol{w}^* = (w_1^*, \ldots, w_p^*)$ are determined by

$$\begin{aligned}&\boldsymbol{w}_* = \mathrm{argmin}_{\boldsymbol{w}^T\mathbf{1}=1}\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w},\\ &\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}\text{: asset covariance matrix.}\end{aligned}$$

In practice, $\boldsymbol{\Sigma}$ is unknown. We first obtain an estimate $\widehat{\boldsymbol{\Sigma}}$ using asset returns $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \in \mathbb{R}^p$ during a period of n days, then we estimate the weights by

$$\widehat{\boldsymbol{w}}_* = \mathrm{argmin}_{\boldsymbol{w}^T\mathbf{1}=1}\boldsymbol{w}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{w}.$$

This optimization has an explicit solution:

$$\widehat{\boldsymbol{w}}_* = (\mathbf{1}^T\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{1})^{-1}(\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{1}). \tag{13}$$

Since what we actually need is $\widehat{\boldsymbol{\Sigma}}^{-1}$, exploring the low-rank plus diagonal dominance structure is a potentially useful strategy.

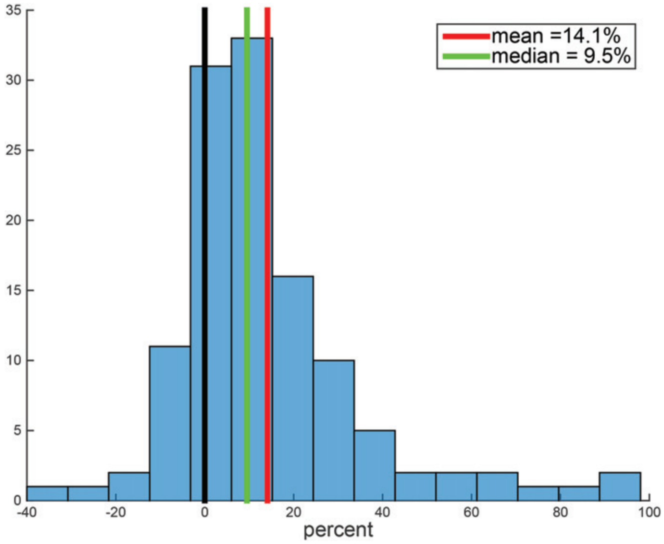

Figure 2. Histogram of ratio of improvement of our method over POET over 120 months.

We compare our method with POET on real data. We collected the daily returns of stocks in S&P 100 index from January 1, 2006 to December 31, 2016. After removing companies that were listed after 2006, there are 80 stocks in total. On the first trading day of each month, we created two portfolios from (13), where $\widehat{\boldsymbol{\Sigma}}$ is estimated using daily returns for the proceeding 12 months ($n = 252$) by our method and by POET, respectively. We set $K = 3$ for both methods. The threshold in POET is chose by cross-validation (we use the default cross-validation procedure in *poet* package). On the last trading day of the same month, we measure the actual risk of each portfolio by

$$R(\widehat{\boldsymbol{w}}_*) = \frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{y}_t^T\widehat{\boldsymbol{w}}^*)^2,$$

where T is the number of trading days in this month ($T = 21$ for most months) and $\boldsymbol{y}_t \in \mathbb{R}^{80}$ contains the stock returns on day t of the month. Define $r = (R_{\text{poet}} - R_{\text{ddpca}})/R_{\text{ddpca}}$; note that a positive r indicates that the portfolio created using our method is superior to that of POET. Figure 2 displays the histogram of r over 120 months in our data range. It suggests that our method improves POET by 9.5% on average and 14.7% in the median.

2.2. Application to Linear Discriminant Analysis

In binary classification, given feature vectors $X_1, \dots, X_n \in \mathbb{R}^p$ and training labels $\ell_1, \dots, \ell_n \in \{1, 2\}$, we aim to construct a linear classifier. In the classical regime where p is fixed as the training sample size grows, Fisher's LDA is an effective linear classifier. In the modern high-dimensional settings where $p \gg n$, it has been well understood that feature screening is necessary before one applies Fisher's LDA (Donoho and Jin 2008; Fan and Fan 2008), and that it is desirable to plug in a good estimate of the inverse covariance matrix that explores structural assumptions (Cai, Liu, and Luo 2011). Recently, Fan, Jin, and Yao (2013) proposed a linear classifier that uses an estimate of inverse covariance matrix in both the screening step and LDA step, and they showed that this classifier is rate-optimal under a multivariate normal model with even extremely weak signal strength. This classifier was later applied to several large real classification problems with superior results (Huang, Jin, and Yao 2016). We shall combine our covariance matrix estimator with this classifier to see whether exploring the low-rank plus diagonal-dominance structure is helpful.

Given an estimate $\widehat{\boldsymbol{\Omega}}$ of the inverse covariance matrix and a threshold $t > 0$, the classifier has four steps (Huang, Jin, and Yao 2016):

1. Calculate the feature-wise t-score: For $1 \le j \le p$, let $Z(j) = [\bar{X}_1(j) - \bar{X}_2(j)]/(n \cdot s_j)$, where $\bar{X}_1(j)$ and $\bar{X}_2(j)$ are the within-class sample means of feature j and $s_j > 0$ is the pooled standard deviation of feature j. Write $Z = (Z(1), \dots, Z(p))^T$.
2. Apply the innovated transformation (Fan, Jin, and Yao 2013) to get $\tilde{Z} = \widehat{\boldsymbol{\Omega}}Z$.
3. Feature-wise thresholding: For $1 \le j \le p$, let $w(j) = \mathrm{sgn}(\tilde{Z}(j)) \cdot 1\{|\tilde{Z}(j)| \ge t\}$. Write $w = (w(1), w(2), \dots, w(p))^T$.
4. Classification by LDA. Given a test feature vector $\widetilde{X} \in \mathbb{R}^p$, for $1 \le j \le p$, normalize $\widetilde{X}(j)$ to $\widetilde{X}^*(j) = [\widetilde{X}(j) - \frac{1}{2}(\bar{X}_1(j) + \bar{X}_2(j))]/s_j$, where $(\bar{X}_1(j), \bar{X}_2(j), s_j)$ are the same as in Step 1. Write $\widetilde{X}^* = (\widetilde{X}^*(1), \dots, \widetilde{X}^*(p))^T$. We classify the test sample to class 1 if $w^T\widehat{\boldsymbol{\Omega}}\widetilde{X}^* > 0$ and to class 2 otherwise.

In this classifier, the matrix $\widehat{\boldsymbol{\Omega}}$ plays two roles: First, it is used in the innovated transformation, so different $\widehat{\boldsymbol{\Omega}}$ leads to different feature rankings. Second, it is used in the LDA step, so $\widehat{\boldsymbol{\Omega}}$ also affects the classification boundary.

We compare the classification performance of plugging in three versions of $\widehat{\boldsymbol{\Omega}}$: The first is $\widehat{\boldsymbol{\Sigma}}_{\text{ddpca}}^{-1}$, the second is $\widehat{\boldsymbol{\Sigma}}_{\text{poet}}^{-1}$, and the last is $[\text{diag}(\boldsymbol{S})]^{-1}$, where $\boldsymbol{S}$ is the sample covariance matrix. We note that the last approach is indeed the method FAIR (Fan and Fan 2008). The above classifier also requires a threshold $t > 0$. To minimize the effects of selecting t, for each $1 \le k \le p$, we set the threshold such that k features are retained and record the classification error. This generates an error curve for each method as k ranges from 1 to p.

We consider two datasets: the lung cancer dataset (Gordon et al. 2002) and the breast cancer dataset (Wang et al. 2005). They were downloaded from *http://blog.nus.edu.sg/staww/softwarecode/*. For both datasets, we conducted a preprocessing by ranking all features by the feature-wise t-score and retaining p_0 top-ranked features, where p_0 is a number that is for sure larger than the true number of useful features (but $p_0 \ll p$).

Dataset	Sample size	Dimension	p_0
Lung cancer	181	12,533	100
Breast cancer	276	22,215	1000

The lung cancer dataset was analyzed in various articles (Tibshirani et al. 2002; Fan and Fan 2008). The estimated number of useful features by these methods is around 30, so we confidently set $p_0 = 100$. The breast cancer dataset is a more difficult one and requires a lot more retained features. Jin and Wang (2016) analyzed the dataset under a clustering framework and suggested that the number of useful features is 728, so we set $p_0 = 1000$. We also tried other choices of p_0 (e.g., $p_0 = 200$ for lung cancer data and $p_0 = 2000$ for breast cancer data), and the results are similar.

We evaluate the classification performance by a 5-fold cross-validation procedure with stratified sampling. In detail, we randomly divide samples from class 1 into five folds and do the same to samples from class 2; we then recombine them to five folds, such that the fraction of class 1 is the same across all folds. Next, we successively leave out each fold, train the classifier on remaining samples, and compute the test error on leave-out samples. The misclassification error reported is the average over 5 folds.

Figure 3 displays the results on lung cancer dataset. POET and DD-PCA have a tuning integer K, and we tried $K \in \{1, 2, 3\}$. The results suggest that, as long as more than 10 features are retained, the classifier powered by DD-PCA uniformly outperforms the other two. Especially, for $K \in \{2, 3\}$, the error keeps as low as 1/181 once the number of retained features exceeds 60. The performance of POET is slightly worse than FAIR for $K \in \{1, 2\}$, and slightly better for $K = 3$. We emphasize that the estimated inverse covariance matrix affects both the feature ranking and the LDA; therefore, even when the number of retained features is the same, the actual retained features are different across different methods.

Figure 4 displays the results on breast cancer dataset. For both POET and DD-PCA, $K \in \{4, 5\}$ is favored to $K = 3$. When $K = 4$, as the number of retained features is in the interval of $[500, 700]$, DD-PCA achieves the smallest error of 114/276. In all three panels, the lowest attainable error of DD-PCA is smaller than those of POET and FAIR.

3. Detecting Sparse Mean Effects by DD-PCA

The global detection is a problem of great interest in multiple testing (Simes 1986; Donoho and Jin 2004; Wu et al. 2011). Let $X_1, \ldots, X_p \in \mathbb{R}$ be the z-scores of p tests, where p is presumably large. We assume

$$X \sim \mathcal{N}_p(\mu, \boldsymbol{\Sigma}), \tag{14}$$

where μ contains the true effects of these tests and $\boldsymbol{\Sigma}$ captures the dependence among the z-scores. We are interested in testing

$$H_0 : \mu = \mathbf{0}, \quad \text{versus} \quad H_1 : \mu \neq \mathbf{0}, \text{ and } \mu \text{ is sparse.} \tag{15}$$

When $\boldsymbol{\Sigma}$ is a diagonal matrix, this problem has been studied extensively in the literature. Various tests were proposed, such as the χ^2 test (or Hotelling's T^2 test), maximum entry test (or minimum p-value test), HC test (Donoho and Jin 2004), Berk–Jones test (Jager and Wellner 2007), etc. The HC test achieves the theoretically optimal detection boundary when the nonzero effects in μ are rare and weak (Donoho and Jin 2004) and has gained increasing popularity in real applications (Donoho and Jin 2015). However, there is limited understanding of how to use the HC test beyond a diagonal $\boldsymbol{\Sigma}$. When $\boldsymbol{\Sigma}$ is heavily non-sparse, a brute-forth application of the orthodox HC test leads to suboptimal performance (Hall and Jin 2008). A satisfactory answer is only available when $\boldsymbol{\Sigma}^{-1}$ is row-wise sparse. Hall and Jin (2010) introduced the "innovated transformation" on data, $X \mapsto \widehat{\boldsymbol{\Omega}}X$, where $\widehat{\boldsymbol{\Omega}}$ is an estimator of $\boldsymbol{\Sigma}^{-1}$. They showed that an application of the HC test on the post-transformation data leads to theoretically optimal testing performance.

Motivated by the popularity of adopting factor covariance structures in multiple testing (Leek and Storey 2008; Fan, Han, and Gu 2012), we consider the global testing problem (14) and (15) by assuming that $\boldsymbol{\Sigma}$ has a low rank plus diagonal dominance structure as in (1) and (2). Denote by $\sum_{k=1}^{K} \nu_k \eta_k \eta_k^T$ the eigen-decomposition of $\boldsymbol{L}$. Equivalently, we model that

$$X = \mu + \sum_{k=1}^{K} w_k \eta_k + z, \tag{16}$$

$$w_k \sim \mathcal{N}(0, \nu_k), \; z \sim \mathcal{N}_p(0, \boldsymbol{A}), \; w_1, \ldots, w_K, z \text{ are independent.}$$

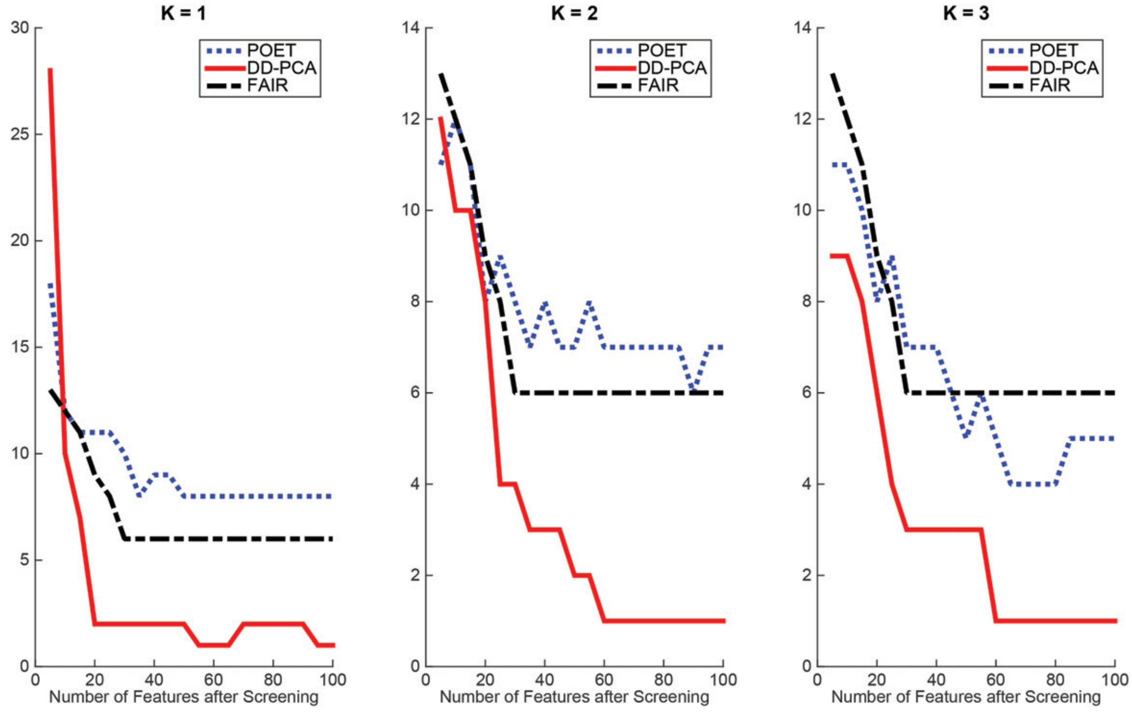

Figure 3. Misclassification errors on lung cancer data ($n = 181$).

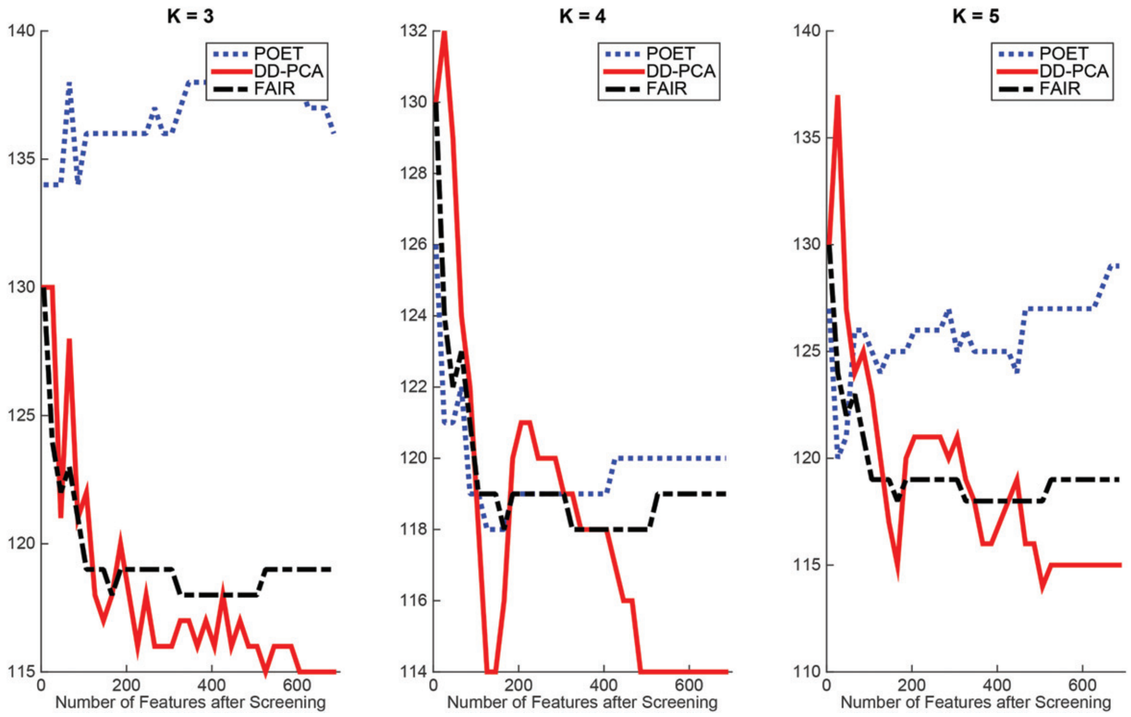

Figure 4. Results for breast cancer data ($n = 276$)

Here, $w_1, \dots, w_K$ are latent variables that account for most of the heavy dependence among test statistics. Under this model, $\mathbf{\Sigma}$ is heavily nonsparse, so the orthodox HC test performs unsatisfactorily (Hall and Jin 2008). At the same time, $\mathbf{\Sigma}^{-1}$ may not be row-wise sparse, so the innovated HC test (Hall and Jin 2010) is not necessarily a good choice either. How to adapt the HC test to factor covariance structures is still largely unclear. We use DD-PCA (and its variants such as one-step DD-PCA) to develop two modifications of the HC test. Both tests significantly outperform the orthodox HC and innovated HC, when the factor covariance structure holds.

Without loss of generality, we assume an estimate of $\mathbf{\Sigma}$ is available, denoted as $\widehat{\mathbf{\Sigma}}$. Note that it is common that a z-score X_j is computed from a number of repeated observations. Suppose we observe iid samples $X_1, X_2, \dots, X_n$ and obtain the z-scores as $X = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$, then the sample covariance matrix of X_i's can be used as $\widehat{\mathbf{\Sigma}}$.

We begin with describing the orthodox HC. Let $\pi_1, \pi_2, \dots, \pi_p$ be the marginal p-values computed from $X_j \sim \mathcal{N}(0, \widehat{\Sigma}_{jj})$, where $\widehat{\Sigma}_{jj}$ is the jth diagonal of $\widehat{\mathbf{\Sigma}}$. Sort the p-values in the descending order and denote by $\pi_{(k)}$ the kth smallest p-value, $1 \le k \le p$. The OHC test statistic is

$$\mathrm{HC}_p^* = \max_{1 \le j \le p/2} \mathrm{HC}_{p,j}, \qquad \text{where} \qquad (17)$$

$$\mathrm{HC}_{p,j} = \frac{\sqrt{p}[(j/p) - \pi_{(j)}]}{\sqrt{\pi_{(j)}(1 - \pi_{(j)})}}, \qquad 1 \le j \le p.$$

The null distribution of HC_p^* is often approximated by a Gumbel distribution or by simulating the null data (Donoho and Jin 2015). Although OHC was proposed for the case of a diagonal $\mathbf{\Sigma}$, we can treat it as a blackbox procedure (we actually know what is happening in the "blackbox," but we have no intention to interfere): It takes as input the p-value for each individual test, and outputs a test statistic for the global null hypothesis which is an aggregation of all individual p-values. Now, if we feed this "blackbox" with a different set of individual p-values, it will output a different test statistic. Following this strategy, we modify OHC by constructing individual p-values using $\widehat{\mathbf{\Sigma}}$, in hopes of borrowing strength from each other (Figure 5).

The first test we propose is *IHC-DD test*, where "IHC" stands for innovated HC and "DD" stands for DD-PCA. As the name has suggested, this test is built on top of the innovated HC test—we replace $\widehat{\mathbf{\Omega}}$ in IHC by the DD-PCA estimator of $\mathbf{\Sigma}^{-1}$ introduced in Section 2.

- Take $\widehat{\mathbf{\Sigma}}$ as the input to the one-step DD-PCA algorithm in Section 1 and let $(\widehat{\boldsymbol{L}}, \widehat{\boldsymbol{A}})$ be the output. Let $\widehat{\mathbf{\Sigma}}_{\text{ddpca}} = \widehat{\boldsymbol{L}} + \widehat{\boldsymbol{A}}$.
- Obtain $\widetilde{X} = \widehat{\mathbf{\Sigma}}_{\text{ddpca}}^{-1} X$. Compute the individual p-values $\widetilde{\pi}_j$ from the null distribution of $\widetilde{X}_j \sim \mathcal{N}(0, \widehat{\Omega}_{jj})$, where $\widehat{\Omega}_{jj}$ is the jth diagonal of $\widehat{\mathbf{\Sigma}}_{\text{ddpca}}^{-1}$.
- Input the p-values $\widetilde{\pi}_1, \dots, \widetilde{\pi}_p$ to the OHC procedure (17) to get a test statistic.

The individual p-values fed to OHC are different from before: Each $\widetilde{\pi}_j$ borrows information from z-scores of other tests, taking advantage of the dependence between tests. The IHC-DD test can be viewed as another application of the DD-PCA covariance estimator proposed in Section 2, where we simply plug the estimated $\mathbf{\Sigma}^{-1}$ into the existing IHC test.

The second test we propose is more customized to the factor covariance structure. We call it *DD-HC* test, to differentiate it from the test above. By (16),

$$X - \sum_{k=1}^{K} w_k \eta_k \sim \mathcal{N}_p(\mu, \boldsymbol{A}).$$

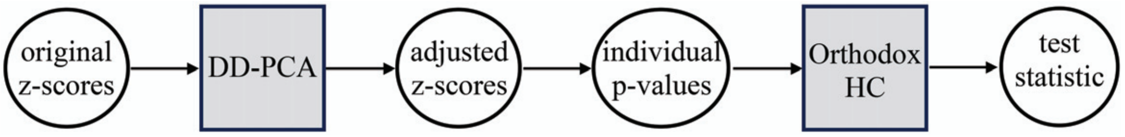

Figure 5. Illustration of the use of DD-PCA to modify the orthodox HC.

Provided with estimates of w_k's, η_k's, and $\boldsymbol{A}$, we can use $X^* = X - \sum_{k=1}^K \widehat{w}_k \widehat{\eta}_k$ as the new z-scores for all individual tests, and we can compute the individual p-values from the null distribution of $X_j^* \sim \mathcal{N}(0, \widehat{A}_{jj})$, where $\widehat{A}_{jj}$ is the jth diagonal of $\widehat{\boldsymbol{A}}$. Let's discuss how to estimate $(w_1, \ldots, w_K, \eta_1, \ldots, \eta_K, \boldsymbol{A})$ from DD-PCA. Let $(\widehat{\boldsymbol{L}}, \widehat{\boldsymbol{A}})$ be the output of DD-PCA with $\widehat{\boldsymbol{\Sigma}}$ as the input (we will discuss below which DD-PCA algorithm to use). Recall that in model (16), η_k's are eigenvectors of $\boldsymbol{L}$. This motivates us to estimate η_k's by the leading eigenvectors of $\widehat{\boldsymbol{L}}$. Once we have $\widehat{\eta}_1, \ldots, \widehat{\eta}_K$, we can approximate model (16) by

$$X = \mu + \sum_{k=1}^K w_k \widehat{\eta}_k + z.$$

This is indeed a linear regression model with K covariates and p observations. The regression coefficients are $w_1, \ldots, w_K$, and each observation has an individual intercept μ_j. Since μ is a sparse mean vector, the majority of μ_j's are zero; we thus estimate $w_1, \ldots, w_K$ by a robust regression via minimizing $\|X - \sum_{k=1}^K w_k \widehat{\eta}_k\|_1$, with respect to w_k's. Last, we discuss the estimation of $\boldsymbol{A}$. A straightforward idea is to use $\widehat{\boldsymbol{A}}$ output by DD-PCA. However, we do not recommend this approach. The estimator of $\boldsymbol{A}$ shall be used to approximate the null distribution of $X_j^* = X_j - \sum_{k=1}^K \widehat{w}_k \widehat{\eta}_k(j)$, to compute individual p-values. So, our goal is not to estimate $\boldsymbol{A}$ accurately but to provide an accurate estimate of the variance of X_j^*. We have to take into account the plug-in effect of $\widehat{w}_k$ and $\widehat{\eta}_k$. Since X_j^* is the residual of fitting a linear regression, a more reasonable choice is to use the diagonals of $\widehat{\boldsymbol{\Sigma}} - \widehat{\boldsymbol{L}}$ as estimates of the variances of X_j^*'s. We summarize the procedure as follows:

- Take $\widehat{\boldsymbol{\Sigma}}$ as the input to a DD-PCA algorithm and let $(\widehat{\boldsymbol{L}}, \widehat{\boldsymbol{A}})$ be the output.
- Conduct PCA on $\widehat{\boldsymbol{L}}$, and denote by $\widehat{\eta}_1, \ldots, \widehat{\eta}_K$ the first K eigenvectors of $\widehat{\boldsymbol{L}}$.
- Regress X on $(\widehat{\eta}_1, \ldots, \widehat{\eta}_K)$ using a robust regression: $\min_{w_1, \ldots, w_K} \|X - \sum_{k=1}^K w_k \widehat{\eta}_k\|_1$. Let $(\widehat{w}_1, \ldots, \widehat{w}_K)$ be the estimated coefficients.
- Obtain $X^* = X - \sum_{k=1}^K \widehat{w}_k \widehat{\eta}_k$. Compute the individual p-values π_j^* from the null distribution of $X_j^* \sim \mathcal{N}(0, \widehat{R}_{jj})$, where $\widehat{R}_{jj}$ is the jth diagonal of $\widehat{\boldsymbol{R}} \equiv \widehat{\boldsymbol{\Sigma}} - \widehat{\boldsymbol{L}}$.
- Input the p-values $\pi_1^*, \ldots, \pi_p^*$ to the OHC procedure (17) to get a test statistic.

Since the variance of X_j^* is smaller than the variance of X_j, we expect that feeding to OHC with these new p-values π_j^* will lead to more testing power. Now, let's consider the choice of the DD-PCA algorithm. We shall use the iterative projection algorithm to be introduced in Section 4.2. The one-step DD-PCA in Section 1 is equivalent to running this iterative algorithm with only one iteration. For the sake of constructing the DD-HC test statistic, we need more iterations. By design of the iterative algorithm, as the number of iterations increases, $\|\widehat{\boldsymbol{\Sigma}} - \widehat{\boldsymbol{L}} - \widehat{\boldsymbol{A}}\|_F$ continues to decrease. It indicates that $\widehat{\boldsymbol{R}}$ and $\widehat{\boldsymbol{A}}$ are closer to each other, and so $\widehat{\boldsymbol{R}}$ becomes more diagonal dominant (note that $\widehat{\boldsymbol{A}}$ is forced to be in the diagonal dominant cone). The matrix $\widehat{\boldsymbol{R}}$ approximately captures the dependence structure in X^*. When $\widehat{\boldsymbol{R}}$ is close to being diagonal, it means the dependence among original z-scores has been fully utilized and there is minimal loss by ignoring the dependence among entries of X_j^*.

Remark. We are not the first to consider using a factor covariance structure to adjust p-values in multiple testing. Fan, Han, and Gu (2012) proposed a similar idea, but their $(\widehat{\boldsymbol{L}}, \widehat{\boldsymbol{A}})$ are from the classical PCA. Our innovation is 2-fold. First, we are the first to incorporate a factor covariance structure in testing against the global null hypothesis. The main focus of Fan, Han, and Gu (2012) is on estimating the false discovery proportion (FDP). Although they briefly mentioned the idea of computing factor-adjusted p-values, they did not use it in their method (their FDP estimator is still based on the original p-values). Second, our method is equipped with the fresh DD-PCA algorithm, in contrast with the classical PCA used in Fan, Han, and Gu (2012).

We investigate the performance of new tests on extensive simulations. Given (n, p, s, τ), let $\mu_j = \tau \cdot 1\{1 \leq j \leq s\}$ for $1 \leq j \leq p$. Generate the matrix $\boldsymbol{\Sigma} = \boldsymbol{F}\boldsymbol{F}^T + \boldsymbol{A}$, where $\boldsymbol{F}$ is a $p \times 2$ matrix whose entries are iid drawn from $\mathcal{N}(0, 1/2)$ and $A_{i,j} = 0.5^{|i-j|}$ for $1 \leq i, j \leq p$. In the null and alternative hypothesis, we generate $X_1, X_2, \ldots, X_n$ iid from $\mathcal{N}_p(0, \boldsymbol{\Sigma})$ and $\mathcal{N}_p(\mu, \boldsymbol{\Sigma})$, respectively. The vector of z-scores is $X = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$, and $\widehat{\boldsymbol{\Sigma}}$ is chosen as the sample covariance matrix of X_i's. For each test, we record the *ideal testing error*, defined as the sum of Type I and Type II errors with the optimal cut-off value for this test (computed via 1000 repetitions). We fix $n = 50$ and let (p, s, τ) take different values, where s controls the sparsity level and τ controls the signal strength.

We compare our test with the χ^2-test (test statistic: $\|X\|^2$), maximum test (test statistic: $\max_{1 \leq j \leq p} |X_j|$), the HC test, and the innovated HC test ($\widehat{\boldsymbol{\Omega}}$ is taken as the generalized inverse of sample covariance matrix). The results are displayed in Figure 6. We have several observations. First, the two proposed tests, IHC-DD and DD-HC, significantly outperform the other tests. Especially, the DD-HC test yields the lowest error in almost all settings. A possible reason is that DD-HC is customized to the factor covariance structure. Second, the IHC-DD test is much better than the IHC test. Since IHC-DD is a variant of IHC by plugging in $\widehat{\boldsymbol{\Sigma}}_{\text{ddpca}}^{-1}$, for a fair comparison, we include two other variants of IHC by plugging $\widehat{\boldsymbol{\Sigma}}_{\text{poet}}^{-1}$ and $\widehat{\boldsymbol{\Sigma}}_{\text{Ind}}^{-1}$, respectively, where $\widehat{\boldsymbol{\Sigma}}_{\text{poet}}$ is the POET estimator and $\widehat{\boldsymbol{\Sigma}}_{\text{Ind}}$ is a special case of POET by keeping only the diagonal entries

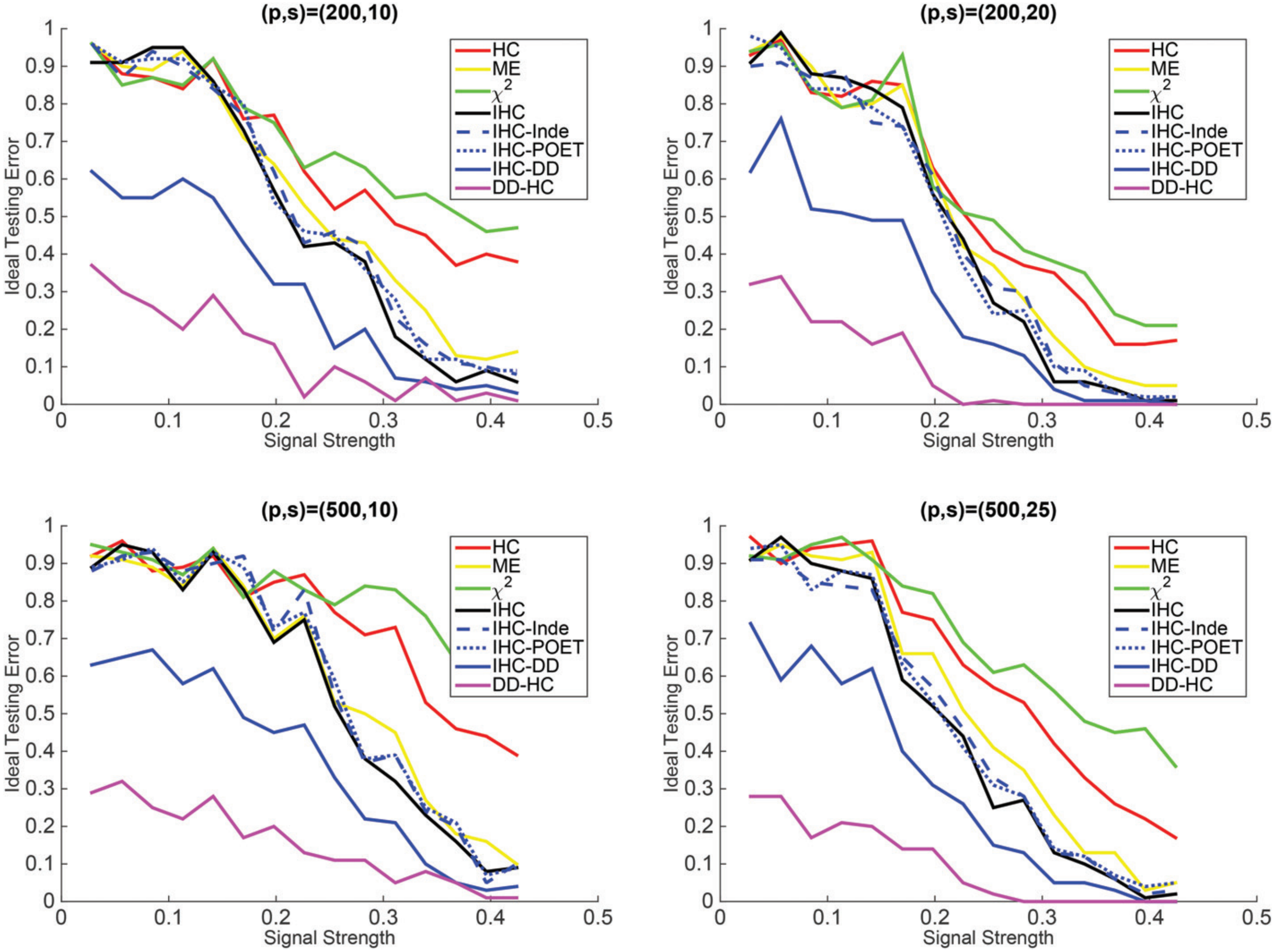

Figure 6. Ideal testing error (with the best cut-off value of the test statistics).

after PCA. Interestingly, these two variants behave similarly as IHC, without much improvement. These observations suggest that the IHC idea is still quite powerful for factor covariance structure, except that we need to plug in a good estimate of $\boldsymbol{\Sigma}^{-1}$; they also reconfirm that the DD-PCA covariance estimator does a good job in estimating $\boldsymbol{\Sigma}^{-1}$. Last, the χ^2-test, orthodox HC, and maximum test perform unsatisfactorily. Since $\boldsymbol{\Sigma}$ is heavily nonsparse, these tests lose power due to not exploring the covariance structure; comparably, the maximum test is less affected.

4. Optimization for DD-PCA

This section studies the optimization problem for DD-PCA. Section 4.1 proposes a three-block alternating direction method of multipliers (ADMM) with provable theoretical guarantees to solve the convex relaxation of DD-PCA. Section 4.2 proposes an iterative projection algorithm to directly solve the nonconvex optimization of DD-PCA. Section 4.3 gives a comparison between two approaches. Both approaches are implemented in the R package "ddpca".

Before proceeding, we introduce the efficient projection onto $\mathcal{SDD}_c^+ = \mathcal{S} \cap \mathcal{DD}_c^+$, the set of "symmetric c-diagonally dominant" matrices, where $\mathcal{S}$ is the set of symmetric matrices and $\mathcal{DD}_c^+$ is the set of c-diagonally dominant matrices with nonnegative diagonal entries:

$$\mathcal{DD}_c^+ = \{\boldsymbol{A} = (a_{ij})_{p\times p} : a_{jj} \ge c \sum_{i:i\neq j} |a_{ji}| \text{ for all } j\}. \qquad (18)$$

We include the details about the efficient projection onto the set of "symmetric c-diagonally dominant" matrices in Appendix A of the online supplementary note. It is not difficult to see that both $\mathcal{DD}_c^+$ and $\mathcal{SDD}_c^+$ are closed and convex polyhedral cones. To solve DD-PCA, we shall obtain the (Euclidean) projection of a matrix $\boldsymbol{A}$ onto the convex cone $\mathcal{SDD}_c^+$ or $\mathcal{DD}_c^+$, denoted by $\mathcal{P}_{\mathcal{SDD}_c^+}(\boldsymbol{A})$ or $\mathcal{P}_{\mathcal{DD}_c^+}(\boldsymbol{A})$. As summarized in Appendix A, the Mendoza–Raydan–Tarazaga (MRT) algorithm computes the efficient projection $\mathcal{P}_{\mathcal{DD}_c^+}(\boldsymbol{A})$. Following Theorem 2.1 of Mendoza, Raydan, and Tarazaga (1998), we have the convergence guarantee that $\boldsymbol{X}$ obtained by MRT algorithm is the unique projection of $\boldsymbol{A}$ onto $\mathcal{DD}_c^+$. The computational complexity of MRT algorithm is $O(p^2 \log(p))$.

4.1. Convex Relaxation and ADMM

This subsection solves the convex relaxation of (4) by replacing nonconvex rank constraints with convex nuclear norm constraints. To be specific, we consider the convex optimization:

$$\min_{(\boldsymbol{L},\boldsymbol{A})} \frac{1}{2}\|\boldsymbol{S}-\boldsymbol{L}-\boldsymbol{A}\|_F^2+\lambda\|\boldsymbol{L}\|_* \quad \text{subject to} \quad \boldsymbol{A} \in \mathcal{SDD}_c^+, \tag{19}$$

where $\|\cdot\|_*$ denotes the matrix nuclear norm and λ is a tuning parameter to strike a balance between the approximation error and the low-rank. A large λ would encourage the solution $\widehat{\boldsymbol{L}}$ to be low-rank, whereas a smaller λ would lead to relatively smaller approximation error but higher rank in $\widehat{\boldsymbol{L}}$. We will present the proposed two-block ADMM for solving the convex relaxation of exact DD-PCA in Appendix B of the online supplementary note.

We introduce a new variable $\boldsymbol{E}$ and rewrite the optimization problem as follows:

$$\min_{(\boldsymbol{L},\boldsymbol{A},\boldsymbol{E})} \frac{1}{2}\|\boldsymbol{E}\|_F^2+\lambda\|\boldsymbol{L}\|_*+\mathcal{I}_{\boldsymbol{A}\in\mathcal{SDD}_c^+} \quad \text{subject to} \quad \boldsymbol{L}+\boldsymbol{A}+\boldsymbol{E}=\boldsymbol{S}.$$

Here, $\mathcal{I}_C$ is the indicator function which equals to 0 if condition C is satisfied, and equals to ∞ otherwise. The objective function would be separable in three blocks, subject to an equality constraint. Now, we define the following augmented Lagrange function:

$$\begin{aligned}\mathcal{L}_\rho(\boldsymbol{L},\boldsymbol{A},\boldsymbol{E},\boldsymbol{\Lambda}) = {} & \frac{1}{2}\|\boldsymbol{E}\|_F^2+\lambda\|\boldsymbol{L}\|_*+\mathcal{I}_{\boldsymbol{A}\in\mathcal{SDD}_c^+} \\ & +\frac{\rho}{2}\|\boldsymbol{L}+\boldsymbol{A}+\boldsymbol{E}-\boldsymbol{S}\|_F^2+\langle\boldsymbol{\Lambda},\boldsymbol{L}+\boldsymbol{A}+\boldsymbol{E}-\boldsymbol{S}\rangle,\end{aligned}$$

where $\boldsymbol{\Lambda}$ is the Lagrange multiplier associated with the equality constraint, and ρ is a given penalty parameter. The proposed three-block ADMM proceeds as follows till convergence:

$$\begin{aligned}
\boldsymbol{L}\text{ step}: \ \boldsymbol{L}^{(t)} &= \arg\min_{\boldsymbol{L}} \mathcal{L}_\rho(\boldsymbol{L},\boldsymbol{A}^{(t-1)},\boldsymbol{E}^{(t-1)},\boldsymbol{\Lambda}^{(t-1)}),\\
\boldsymbol{A}\text{ step}: \ \boldsymbol{A}^{(t)} &= \arg\min_{\boldsymbol{A}} \mathcal{L}_\rho(\boldsymbol{L}^{(t)},\boldsymbol{A},\boldsymbol{E}^{(t-1)},\boldsymbol{\Lambda}^{(t-1)}),\\
\boldsymbol{E}\text{ step}: \ \boldsymbol{E}^{(t)} &= \arg\min_{\boldsymbol{E}} \mathcal{L}_\rho(\boldsymbol{L}^{(t)},\boldsymbol{A}^{(t)},\boldsymbol{E},\boldsymbol{\Lambda}^{(t-1)}),\\
\boldsymbol{\Lambda}\text{ step}: \ \boldsymbol{\Lambda}^{(t)} &= \boldsymbol{\Lambda}^{(t-1)}+\rho(\boldsymbol{A}^{(t)}+\boldsymbol{L}^{(t)}+\boldsymbol{E}^{(t)}-\boldsymbol{S}).
\end{aligned}$$

Each subproblem can be efficiently solved. In the $\boldsymbol{L}$ step, we solve $\boldsymbol{L}^{(t)}$ from

$$\begin{aligned}
&\min\nolimits_{\boldsymbol{L}}\ \lambda\|\boldsymbol{L}\|_*+\frac{\rho}{2}\|\boldsymbol{L}+\boldsymbol{A}^{(t-1)}+\boldsymbol{E}^{(t-1)}-\boldsymbol{S}\|_F^2\\
&\qquad +\langle\boldsymbol{\Lambda}^{(t-1)},\boldsymbol{L}+\boldsymbol{A}^{(t-1)}+\boldsymbol{E}^{(t-1)}-\boldsymbol{S}\rangle\\
\Longleftrightarrow\ &\min\nolimits_{\boldsymbol{L}}\ \frac{1}{2}\|\boldsymbol{L}+\boldsymbol{A}^{(t-1)}+\boldsymbol{E}^{(t-1)}-\boldsymbol{S}+\rho^{-1}\boldsymbol{\Lambda}^{(t-1)}\|_F^2\\
&\qquad +\rho^{-1}\lambda\|\boldsymbol{L}\|_*.
\end{aligned}$$

We have $\boldsymbol{L}^{(t)} = \mathcal{D}_{\rho^{-1}\lambda}\left(\boldsymbol{S}-\boldsymbol{A}^{(t-1)}-\boldsymbol{E}^{(t-1)}-\rho^{-1}\boldsymbol{\Lambda}^{(t-1)}\right)$, where $\mathcal{D}_\tau$ is the singular value thresholding operator $\mathcal{D}_\tau(\boldsymbol{\Omega}) = \boldsymbol{U}s_\tau(\boldsymbol{D})\boldsymbol{V}^T$ for any singular value decomposition $\boldsymbol{\Omega}=\boldsymbol{UDV}^T$, and s_τ denotes the soft-thresholding operator $s_\tau(x) = \operatorname{sgn}(x)\max(|x|-\tau,0)$.

In the $\boldsymbol{A}$ step, we need to obtain the projection on $\mathcal{SDD}_c^+$:

$$\begin{aligned}
\boldsymbol{A}^{(t)} &= \arg\min_{\boldsymbol{A}} \mathcal{I}_{\boldsymbol{A}\in\mathcal{SDD}_c^+}\\
&\quad +\frac{\rho}{2}\left(\|\boldsymbol{A}+\boldsymbol{L}^{(t)}+\boldsymbol{E}^{(t-1)}-\boldsymbol{S}+\rho^{-1}\boldsymbol{\Lambda}^{(t-1)}\|_F^2\right)\\
&= \mathcal{P}_{\mathcal{SDD}_c^+}\left(\boldsymbol{S}-\boldsymbol{L}^{(t)}-\boldsymbol{E}^{(t-1)}-\rho^{-1}\boldsymbol{\Lambda}^{(t-1)}\right).
\end{aligned}$$

To this end, we follow Mendoza, Raydan, and Tarazaga (1998) to use Dykstra's alternating projection algorithm between $\mathcal{DD}_c^+$ and $\mathcal{S}$. The details of this alternating projection are included in Appendix A of the online supplementary note. Alternatively, we may follow the proximal-gradient-based ADMM (Ma, Xue, and Zou 2013) to solve the $\boldsymbol{A}$ step. See Section 4 of Ma, Xue, and Zou (2013) for more details.

In the $\boldsymbol{E}$ step, it is straightforward to solve

$$\begin{aligned}
\boldsymbol{E}^{(t)} &= \arg\min_{\boldsymbol{E}} \frac{1}{2}\|\boldsymbol{E}\|_F^2+\frac{\rho}{2}\left(\|\boldsymbol{E}+\boldsymbol{L}^{(t)}+\boldsymbol{A}^{(t)}-\boldsymbol{S}+\rho^{-1}\boldsymbol{\Lambda}^{(t-1)}\|_F^2\right)\\
&= \arg\min_{\boldsymbol{E}} \left\|\boldsymbol{E}+\frac{\rho}{\rho+1}\left(\boldsymbol{L}^{(t)}+\boldsymbol{A}^{(t)}-\boldsymbol{S}+\rho^{-1}\boldsymbol{\Lambda}^{(t-1)}\right)\right\|_F^2\\
&= \frac{\rho}{\rho+1}\left(\boldsymbol{S}-\boldsymbol{A}^{(t)}-\boldsymbol{L}^{(t)}-\rho^{-1}\boldsymbol{\Lambda}^{(t-1)}\right).
\end{aligned}$$

Hence, the proposed three-block ADMM can be summarized in Algorithm 1.

Algorithm 1. ADMM for solving the convex relaxation of DD-PCA

Given a sample covariance matrix $\boldsymbol{S}$, do:

- Let $\boldsymbol{A}^{(0)}=\boldsymbol{E}^{(0)}=\boldsymbol{\Lambda}^{(0)}=\boldsymbol{0}$.
- For $t=1,2,\ldots$
  - $\boldsymbol{L}^{(t)} = \mathcal{D}_{\rho^{-1}\lambda}\left(\boldsymbol{S}-\boldsymbol{A}^{(t-1)}-\boldsymbol{E}^{(t-1)}-\rho^{-1}\boldsymbol{\Lambda}^{(t-1)}\right)$ where $\mathcal{D}_\tau(\boldsymbol{\Omega})$ is the singular value thresholding operator given by $\mathcal{D}_\tau(\boldsymbol{\Omega}) = \boldsymbol{U}s_\tau(\boldsymbol{D})\boldsymbol{V}^T$ for any singular value decomposition $\boldsymbol{\Omega}=\boldsymbol{UDV}^T$, and s_τ denotes the soft-thresholding operator given by $s_\tau(x)=\operatorname{sgn}(x)\max(|x|-\tau,0)$.
  - $\boldsymbol{A}^{(t)} = \mathcal{P}_{\mathcal{SDD}_c^+}\left(\boldsymbol{S}-\boldsymbol{L}^{(t)}-\boldsymbol{E}^{(t-1)}-\rho^{-1}\boldsymbol{\Lambda}^{(t-1)}\right)$.
  - $\boldsymbol{E}^{(t)} = \frac{\rho}{\rho+1}\left(\boldsymbol{S}-\boldsymbol{A}^{(t)}-\boldsymbol{L}^{(t)}-\rho^{-1}\boldsymbol{\Lambda}^{(t-1)}\right)$.
  - $\boldsymbol{\Lambda}^{(t)} = \boldsymbol{\Lambda}^{(t-1)}+\rho\left(\boldsymbol{A}^{(t)}+\boldsymbol{L}^{(t)}+\boldsymbol{E}^{(t)}-\boldsymbol{S}\right)$.
- Stop if the convergence criterion is met.

Although three-block ADMM does not necessarily converge in general (Chen et al. 2016), DD-PCA belongs to a class of regularized least squares decomposition problem. For this class of regularized problems, the global convergence of the proposed three-block ADMM is always guaranteed such that any cluster point of the iterated solutions is an optimal primal and dual pair of DD-PCA (see Lin, Ma, and Zhang 2018, Theorem 3.2).

4.2. An Iterative Projection Algorithm

In the sequel, we introduce an iterative projection algorithm that directly tackles the nonconvex optimization in DD-PCA. The key observation is that we attempt to find a matrix $\boldsymbol{L}^*$ in the set $\mathcal{L}_K = \{\boldsymbol{L} : \operatorname{rank}(\boldsymbol{L}) = K\}$ that is closest to the set $\mathcal{M}_{\boldsymbol{S}} = \{\boldsymbol{S}-\boldsymbol{A} : \boldsymbol{A} \in \mathcal{SDD}_c^+\}$. Inspired by Netrapalli et al. (2014), a natural approach would be to iteratively project $(\boldsymbol{S}-\boldsymbol{L})$ onto $\mathcal{SDD}_c^+$ to update $\boldsymbol{A}$ and then to project $(\boldsymbol{S}-\boldsymbol{A})$ onto $\mathcal{L}_K$ to update $\boldsymbol{L}$. To reduce the computational cost, we replace the projection onto $\mathcal{SDD}_c^+$ by the projection onto $\mathcal{DD}_c^+$, followed by symmetrization. Algorithm 2 summarizes the details.

Algorithm 2. Iterative projection algorithm for solving the DD-PCA

Given a sample covariance matrix $\boldsymbol{S}$ and integer k, do:

- Let $\boldsymbol{A}^{(0)} = \boldsymbol{0}$.
- For $t = 1, 2, \ldots$
  - $\boldsymbol{L}^{(t)} = \mathcal{P}_{\mathcal{L}_K}(\boldsymbol{S} - \boldsymbol{A}^{(t-1)})$
  - $\tilde{\boldsymbol{A}}^{(t)} = \mathcal{P}_{\mathcal{DD}_c^+}(\boldsymbol{S} - \boldsymbol{L}^{(t)})$.
  - $\boldsymbol{A}^{(t)} = \left(\tilde{\boldsymbol{A}}^{(t)} + (\tilde{\boldsymbol{A}}^{(t)})^T\right)/2$.
- Stop if the convergence criterion is met.

In Algorithm 2, we need to calculate $\mathcal{P}_{\mathcal{L}_K}$ and $\mathcal{P}_{\mathcal{DD}_c^+}$. The calculation of $\mathcal{P}_{\mathcal{DD}_c^+}$ is given in Appendix A. The calculation of $\mathcal{P}_{\mathcal{L}_K}$ is given as follows: for any symmetric matrix $\boldsymbol{A}$, we write its eigenvalue decomposition as $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T$ where $\boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_p\}$ with $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_p|$. Hence, the best rank-$K$ approximation is given by $\mathcal{P}_{\mathcal{L}_K}(\boldsymbol{A}) = \boldsymbol{Q}_K\boldsymbol{\Lambda}_K\boldsymbol{Q}_K^T$ where $\boldsymbol{Q}_K$ contains the first K columns of $\boldsymbol{Q}$ and $\boldsymbol{\Lambda}_K = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_K\}$.

To use Algorithm 2, we need to estimate the rank K if it is unknown. A simple estimate of K is to look at the eigenvalues of $\boldsymbol{S}$ to pick the K such that there is a significant gap in magnitude between the first K eigenvalues and the remaining ones. In Section 5, we investigate the robustness of the iterative projection algorithm to the estimation of K.

4.3. Comparing Convex and Nonconvex Approaches

The convex approaches do not require the knowledge of rank K of the low rank matrix $\boldsymbol{L}$, and the global convergence of the proposed ADMM is guaranteed. However, its convergence rate could be slow. The nonconvex approaches, on the other hand, can be faster in terms of convergence. The per-iteration cost for Algorithm 2 is $O(p^2 \max\{\log(p), K\})$, compared to $O(p^3)$ for Algorithm 1. But the convergence guarantee of Algorithm 2 is an open question.

The rigorous convergence analysis of the iterative projection algorithm is difficult due to the nonconvexity of the set $\mathcal{L}_K$. The existing result (e.g., Drusvyatskiy, Ioffe, and Lewis 2015) proves the local linear convergence of the alternating projections for two closed sets if the two sets intersect *transversally* at the converging point. We conjecture that such condition would hold for most cases in our setting, therefore, the convergence would be guaranteed. In practice, our algorithms are stable and always converge to a valid solution in simulations.

5. Simulation Studies

This section investigates several numerical properties of DD-PCA, including the estimation performance, necessity and robustness, and application to covariance matrix estimation.

5.1. Experiment 1: Exact DD-PCA

We first examine the numerical performance of two-block ADMM (see Appendix B) and the iterative projection algorithm in Algorithm 2 to solve the exact DD-PCA. Fixing (p, K), we first generate a rank-K matrix $\boldsymbol{L} = XX^T$ where X is a $p \times k$ matrix whose entries are iid drawn from $\mathcal{N}(0, 1/p)$. We then generate a matrix $\boldsymbol{A}_0$ with entries sampled iid from $\mathcal{N}(0, 1/p^2)$ and set $\boldsymbol{A} = \boldsymbol{A}_0 + \boldsymbol{A}_0^T + \boldsymbol{D}$, where $\boldsymbol{D}$ is a diagonal matrix whose jth diagonal is equal to $\sum_{i:i \neq j} |A_0(j,i) + A_0(i,j)| - 2A_0(j,j)$ for $1 \leq j \leq p$; it follows that $\boldsymbol{A}$ is a diagonally dominant matrix. We then let $\boldsymbol{S} = \boldsymbol{L} + \boldsymbol{A}$. We consider $p = 500, 1000, 2000$ and fix $K = 0.05 \cdot p$ for each choice of p.

Table 1. Performance of the proposed ADMM in Experiment 1.

Dimension p	rank($\boldsymbol{L}$)	rank($\widehat{\boldsymbol{L}}$)	$\frac{\|\widehat{\boldsymbol{L}}+\widehat{\boldsymbol{A}}-\boldsymbol{S}\|_F}{\|\boldsymbol{S}\|_F}$	$\frac{\|\widehat{\boldsymbol{L}}-\boldsymbol{L}\|_F}{\|\boldsymbol{L}\|_F}$	$\frac{\|\widehat{\boldsymbol{A}}-\boldsymbol{A}\|_F}{\|\boldsymbol{A}\|_F}$
500	25	25	0.008	0.011	0.045
1000	50	50	0.010	0.008	0.034
2000	100	100	0.013	0.006	0.026

First, for two-block ADMM, we use the solution $(\widehat{\boldsymbol{L}}, \widehat{\boldsymbol{A}})$ after 20 iterations. Table 1 displays the comparison between $(\widehat{\boldsymbol{L}}, \widehat{\boldsymbol{A}})$ and the true $(\boldsymbol{L}, \boldsymbol{A})$ based on 20 repetitions. It suggests that $\widehat{\boldsymbol{L}}$ always has the same rank as that of $\boldsymbol{L}$ and that $\widehat{\boldsymbol{L}}$ and $\widehat{\boldsymbol{A}}$ are reasonably close to their respective counterparts. Since the ADMM algorithm is an iterative algorithm, $\widehat{\boldsymbol{L}}+\widehat{\boldsymbol{A}}$ are not exactly equal to $\boldsymbol{S}$, but the two matrices are reasonably close after 20 iterations. The results also suggest that, for a large p, more iterations are needed for the convergence of ADMM.

Next, we look at Algorithm 2. Instead of fixing the maximum number of iterations, we investigate how the solution $(\widehat{\boldsymbol{L}}, \widehat{\boldsymbol{A}})$ evolves over iterations. Note that $\widehat{\boldsymbol{L}}$ is guaranteed to have rank K in all iterations, and $\widehat{\boldsymbol{A}}$ is equal to the projection of $(\boldsymbol{S} - \widehat{\boldsymbol{L}})$ into $\mathcal{DD}^+$ (with symmetrization). We introduce a quantity to measure the diagonal dominance of $(\boldsymbol{S} - \widehat{\boldsymbol{L}})$. For any matrix $p \times p$ matrix $\boldsymbol{B}$, define

$$\zeta(\boldsymbol{B}) = \min_{1 \leq j \leq p} \left\{ b_{jj} - \sum_{1 \leq i \leq p: i \neq j} |b_{ji}| \right\}.$$

It measures how close a matrix is to the diagonally dominant cone. If $\zeta(\boldsymbol{\Sigma} - \widehat{\boldsymbol{L}})$ continues to increase and eventually gets close to zero, then the algorithm converges. The left panel of Figure 7 shows the evolution of $\zeta(\boldsymbol{S} - \widehat{\boldsymbol{L}})$ over iterations, and it suggests that the algorithm converges quickly. The right panel of Figure 7 displays the evolution of the relative approximation error $\|\widehat{\boldsymbol{L}} + \widehat{\boldsymbol{A}} - \boldsymbol{S}\|_F / \|\boldsymbol{S}\|_F$, which also decreases quickly over iterations.

5.2. Experiment 2: Approximate DD-PCA

We investigate the performance of Algorithm 1 (an ADMM algorithm) and Algorithm 2 (an iterative projection algorithm) for approximate DD-PCA. Fixing (p, K) and $\sigma > 0$, we generate a rank K matrix $\boldsymbol{L}$ and a diagonally dominant matrix $\boldsymbol{A}$ in the same way as in Experiment 1. We then generate a $p \times p$ symmetric matrix $\boldsymbol{E}$ whose upper triangular entries are sampled iid from $\mathcal{N}(0, \sigma^2/p)$. Last, let $\boldsymbol{S} = \boldsymbol{L} + \boldsymbol{A} + \boldsymbol{E}$.

First, we study Algorithm 1, which is an ADMM algorithm. Fix $\sigma = 1$. We consider $p = 500, 1000, 2000$, and set $K = 0.05 \cdot p$. The tuning parameter in the algorithm is set as $\lambda = 3$, and we look at the solution $(\widehat{\boldsymbol{L}}, \widehat{\boldsymbol{A}})$ after 50 iterations. The results are displayed in Table 2. For all three settings, the algorithm exactly recovers the true rank of $\boldsymbol{L}$, however, the convergence of $(\widehat{\boldsymbol{L}}, \widehat{\boldsymbol{A}})$ is relatively slow. As we shall see below, the performance of Algorithm 1 is not as good as the iterative projection algorithm—Algorithm 2, but Algorithm 1 is theoretically more tractable.

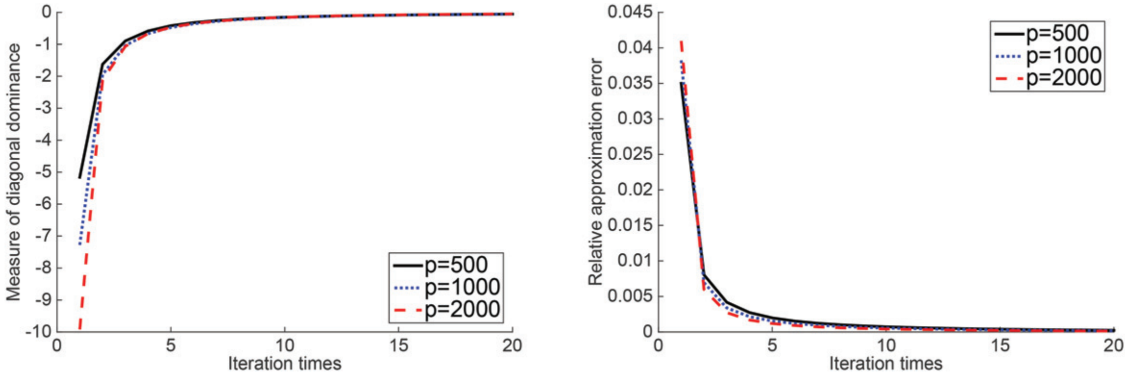

Figure 7. Performance of Algorithm 2 in Experiment 1. The y-axis represents $\zeta(\boldsymbol{\Sigma} - \widehat{\boldsymbol{L}})$ (left panel) and $\|\widehat{\boldsymbol{L}} + \widehat{\boldsymbol{A}} - \boldsymbol{S}\|_F/\|\boldsymbol{S}\|_F$ (right panel).

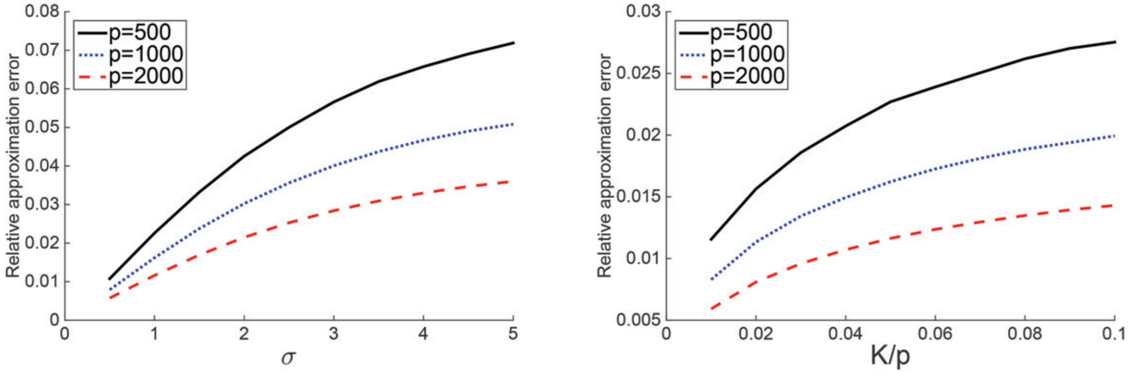

Figure 8. Performance of Algorithm 2 in Experiment 2. The y-axis represents $\|\widehat{\boldsymbol{L}}+\widehat{\boldsymbol{A}}-\boldsymbol{S}\|_F/\|\boldsymbol{S}\|_F$, and the x-axis represents σ (left panel) and K/p (right panel), respectively.

Table 2. Performance of Algorithm 1 in Experiment 2.

Dimension p	rank($\boldsymbol{L}$)	rank($\widehat{\boldsymbol{L}}$)	$\frac{\|\widehat{\boldsymbol{L}}+\widehat{\boldsymbol{A}}-\boldsymbol{S}\|_F}{\|\boldsymbol{S}\|_F}$	$\frac{\|\widehat{\boldsymbol{L}}-\boldsymbol{L}\|_F}{\|\boldsymbol{L}\|_F}$	$\frac{\|\widehat{\boldsymbol{A}}-\boldsymbol{A}\|_F}{\|\boldsymbol{A}\|_F}$
500	25	25	0.264	0.166	0.340
1000	50	50	0.269	0.163	0.286
2000	100	100	0.274	0.160	0.243

Next, we study Algorithm 2, the iterative projection algorithm. In Experiment 1, we have investigated its performance when $\boldsymbol{S}$ has an exact decomposition to the sum of a low-rank matrix and a diagonally dominant matrix. In this experiment, we apply the same algorithm to $\boldsymbol{S}$ which does not have such an exact decomposition. We run the algorithm for 20 iterations and measure the relative approximation error $\|\widehat{\boldsymbol{L}} + \widehat{\boldsymbol{A}} - \boldsymbol{S}\|_F/\|\boldsymbol{S}\|_F$. The results are shown in Figure 8. In the left panel, K/p is fixed as 0.05 and the noise level σ varies from 0.5 to 5. In the right panel, σ is fixed to be 1 and K/p varies from 0.01 to 0.1. For each value of p, the relative approximation error increases, as both σ and K increase. For the same values of σ and K/p, a larger p comes with a smaller relative approximation error. Furthermore, if we compare the results with those in Table 2, Algorithm 2 has a better practical performance than Algorithm 1.

5.3. Experiment 3: Necessity of DD-PCA

If $\boldsymbol{\Sigma}$ truly satisfies the assumption of "low-rank plus diagonal dominance," it is a natural question to know whether one can simply apply PCA and robust PCA (Candès et al. 2011) to get a diagonally dominant $\boldsymbol{A}$. Unfortunately, this is often not the case.

Let us consider applying PCA to a $\boldsymbol{\Sigma}$ which has the decomposition $\boldsymbol{\Sigma} = \boldsymbol{L}_0 + \boldsymbol{A}_0$ such that rank($\boldsymbol{L}_0$) $= K$ and $\boldsymbol{A}_0$ is diagonally dominant. Let λ_k and ξ_k be the kth eigenvalue and eigenvector, respectively, $1 \le k \le p$. We construct $\boldsymbol{L} = \sum_{k=1}^{K} \lambda_k \xi_k \xi_k^T$ and $\boldsymbol{A} = \boldsymbol{\Sigma} - \boldsymbol{L}$. We can only hope $\boldsymbol{A}$ is diagonally dominant when $\boldsymbol{A}$ and $\boldsymbol{A}_0$ are entrywise close to each other, or equivalently, when $\|\boldsymbol{L} - \boldsymbol{L}_0\|_{\max}$ is small ($\|\cdot\|_{\max}$ is the entrywise max norm). However, from the literatures on perturbation analysis of PCA, it requires strong conditions to guarantee that $\|\boldsymbol{L} - \boldsymbol{L}_0\|_{\max}$ is small (Ke and Yang 2017). In particular, when K is moderately large, these conditions may be violated. Similarly, robust PCA cannot produce a diagonally dominant $\boldsymbol{A}$ in general. Therefore, it is necessary to develop new algorithms that are specifically designed for DD-PCA. In Figure 9, we present a numerical example, where the output $\boldsymbol{A}$ from our DD-PCA algorithm is much more "diagonally dominant" than the $\boldsymbol{A}$ constructed from PCA.

5.4. Experiment 4: Robustness to the Misspecification of K

We use the same setup as in Experiment 1 and investigate the performance of Algorithm 2 with a misspecified K. Consider two settings where $(p, K) = (500, 25)$ and $(p, K) = (2000, 100)$, respectively. For each setting, we plug $K = k$ in the algorithm and take the solution after ten iterations. Figure 10 shows the relative difference between $\widehat{\boldsymbol{L}}$ and $\boldsymbol{L}$ for various choices of k. It suggests that as long as $k \ge K$, the performance of the algorithm is very stable. Hence, in practice, we recommend that the users pick a relatively large k when the true K is hard to estimate.

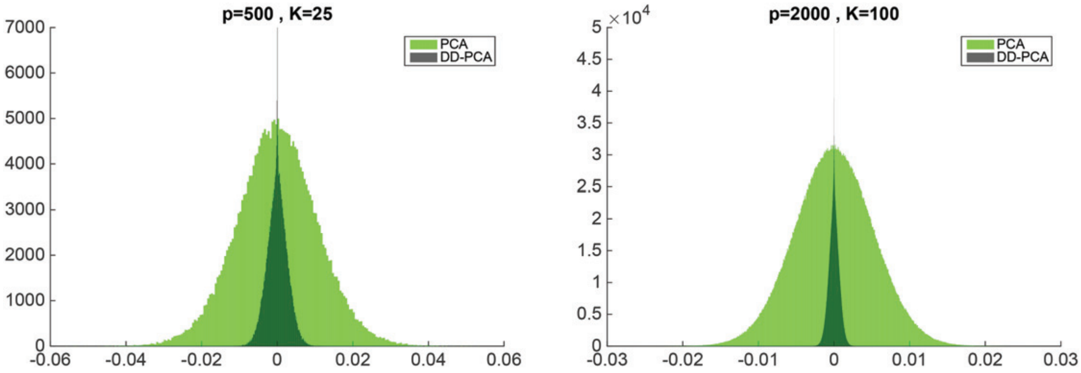

Figure 9. Comparison of the output $\boldsymbol{A}$ from DD-PCA and from PCA, where the histogram of $\{a_{ij}/[a_{ii}a_{jj}]^{1/2} : 1 \leq i \neq j \leq p\}$ is displayed. In both panels, the input $\boldsymbol{\Sigma}$ is generated as in Experiment 1 in Section 5.

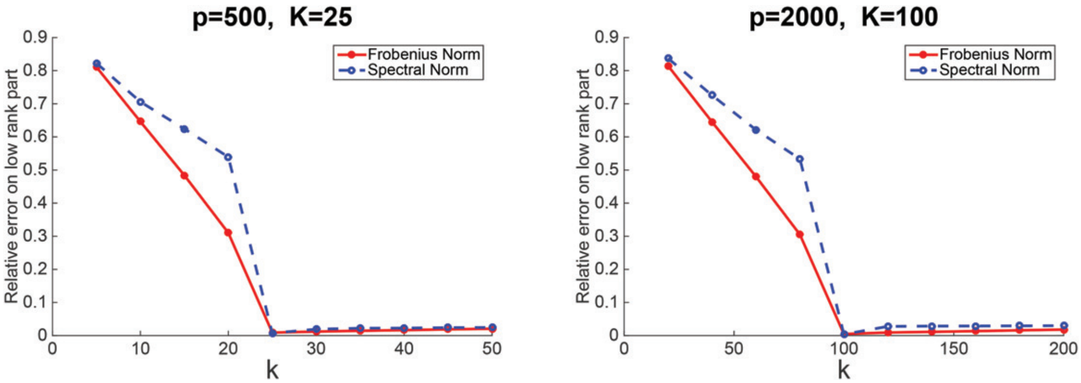

Figure 10. Robustness of Algorithm 2 to a misspecified K. The x-axis is the k used in the algorithm, and the y-axis is $\|\widehat{\boldsymbol{L}} - \boldsymbol{L}\|/\|\boldsymbol{L}\|$, where $\|\cdot\|$ is Frobenius norm or spectral norm.

5.5. Experiment 5: Application to Covariance Matrix Estimation

We expand the numerical study in Section 2 and investigate the performance of DD-PCA on more simulation settings. Given $K = 3$ and $p \in \{100, 300, 500\}$, we generate data in the same way as in the numerical example of Section 2. First, we compare the performance of DD-PCA and POET. For both methods, we use the true $K = 3$. POET has an additional threshold, which we set as the ideal one that minimizes the estimation error (the ideal threshold varies as we change the error measure). The results are contained in Column 6 and Column 10 of Table 3, where, in all settings, DD-PCA has a comparable performance as POET with an ideal threshold, and in some settings, DD-PCA is even better. The ideal threshold for POET is not practically feasible, and it is unclear how to set the threshold in a data-driven fashion; however, DD-PCA is tuning free once K is given. Second, we investigate the performance of DD-PCA when we plug in $K = k$ with $k \in \{1, 2, \ldots, 6\}$; see Table 3. If k is misspecified but $k \geq K$, the estimation errors remain relatively stable; if $k < K$, the performance deteriorates. It suggests that an overshooting of K is better than an undershooting. This is consistent with the observations made by Fan, Liao, and Mincheva (2013).

6. Discussion

The diagonally dominant matrices have been well studied in linear algebra (Feingold and Varga 1962; Barker and Carlson 1975) and optimization (Barlow and Demmel 1990; Mendoza, Raydan, and Tarazaga 1998), motivated by the appealing properties of these matrices for computation. Our work has a very different motivation: We recognize that diagonally dominant (covariance) matrices also have appealing statistical properties, and propose exploring the “low-rank plus diagonal dominance” covariance structure in data analysis. We demonstrate the benefit of exploring such structure in two statistical problems. For covariance matrix estimation, we propose DD-PCA as a new estimator. For testing of the global null hypothesis in multiple testing, we propose IHC-DD and DD-HC as two new tests. These new methods have shown encouraging numerical performance in simulations and real applications, especially when the data have a factor covariance structure.

The above methods rely on the availability of algorithms to decompose any given covariance matrix (approximately) into the sum of a low-rank matrix and a diagonally dominant matrix. To obtain such decomposition is a nonconvex optimization. We propose two algorithms—an ADMM algorithm that solves a convex relaxation, and an iterative projection algorithm that solves the nonconvex problem directly. In comparison, the ADMM algorithm is theoretically more tractable, and the iterative projection algorithm shows very appealing numerical performance.

The study here motivates several interesting future directions, such as the uniqueness of the low-rank plus diagonal dominance decomposition, the statistical error of recovering $\boldsymbol{L}$ and $\boldsymbol{A}$, as well as the convergence of proposed algorithms. We leave to future works.

Table 3. Estimation errors of DD-PCA and its robustness to a misspecified K.

(p, K)	Target	Norm	k						POET*
			1	2	3	4	5	6	$(k = 3)$
(100,3)	$\boldsymbol{\Sigma}_u$	Frobenius	48.26	27.52	3.28	3.46	3.63	3.83	**3.24**
		Spectral	22.51	16.80	**0.80**	1.00	1.08	1.14	0.85
	$\boldsymbol{\Sigma}_u^{-1}$	Frobenius	9.10	7.50	**3.02**	3.17	3.34	3.56	3.65
		Spectral	1.34	1.34	**0.61**	0.72	0.83	0.93	0.64
(300,3)	$\boldsymbol{\Sigma}_u$	Frobenius	95.00	56.96	6.22	6.23	6.26	6.32	**6.04**
		Spectral	32.52	26.04	**0.82**	0.86	0.90	0.93	0.90
	$\boldsymbol{\Sigma}_u^{-1}$	Frobenius	41.50	17.33	**5.68**	5.66	5.66	5.68	6.37
		Spectral	27.75	7.16	0.66	0.66	0.65	0.64	**0.64**
(500,3)	$\boldsymbol{\Sigma}_u$	Frobenius	126.50	75.74	8.38	8.35	8.35	8.35	**7.99**
		Spectral	38.10	30.86	**0.84**	0.87	0.89	0.91	0.95
	$\boldsymbol{\Sigma}_u^{-1}$	Frobenius	25.87	18.19	**7.66**	7.61	7.57	7.55	8.26
		Spectral	9.23	1.76	0.69	0.68	0.68	0.67	**0.64**

*POET is implemented with an ideal threshold.

Our work is related to the literatures of factor models and the literatures of "low-rank plus sparse" matrix decomposition. These two lines of works have found wide applications in many areas. Similarly, the use of DD-PCA is not limited to covariance matrix estimation and multiple testing. We expect that DD-PCA will have applications in classification (Zhu and Hastie 2005; Tong, Feng, and Zhao 2016), clustering (Wang and Zhu 2008; Clarke, Fokoue, and Zhang 2009), dimension reduction (Ma and Zhu 2013; Zou and Xue 2018), and forecasting (Fan, Xue, and Yao 2017).

Supplementary Materials

The online supplementary note includes the details about the efficient projection onto the set of "symmetric c-diagonally dominant" matrices and the proposed algorithm for solving the convex relaxation of exact DD-PCA in Section 4.

Funding

The authors gratefully acknowledge the support of NIH grant P50DA039838 and NSF grants DMS-1505256, DMS-1712958, and DMS-1811552.

References

Barker, G., and Carlson, D. (1975), "Cones of Diagonally Dominant Matrices," *Pacific Journal of Mathematics*, 57, 15–32. [605]

Barlow, J., and Demmel, J. (1990), "Computing Accurate Eigensystems of Scaled Diagonally Dominant Matrices," *SIAM Journal on Numerical Analysis*, 27, 762–791. [605]

Bickel, P. J. (1975), "One-Step Huber Estimates in the Linear Model," *Journal of the American Statistical Association*, 70, 428–434. [593]

Cai, T., and Liu, W. (2011), "Adaptive Thresholding for Sparse Covariance Matrix Estimation," *Journal of the American Statistical Association*, 106, 672–684. [592,594]

Cai, T., Liu, W., and Luo, X. (2011), "A Constrained ℓ_1 Minimization Approach to Sparse Precision Matrix Estimation," *Journal of the American Statistical Association*, 106, 594–607. [597]

Candès, E. J., Li, X., Ma, Y., and Wright, J. (2011), "Robust Principal Component Analysis?," *Journal of the Association for Computing Machinery*, 58, 1–37. [592,604]

Chamberlain, G., and Rothschild, M. (1983), "Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets," *Econometrica*, 51, 1281–1304. [592]

Chen, C., He, B., Ye, Y., and Yuan, X. (2016), "The Direct Extension of ADMM for Multi-Block Convex Minimization Problems Is Not Necessarily Convergent," *Mathematical Programming*, 155, 57–79. [602]

Clarke, B., Fokoue, E., and Zhang, H. H. (2009), *Principles and Theory for Data Mining and Machine Learning*, New York: Springer. [606]

Donoho, D., and Jin, J. (2004), "Higher Criticism for Detecting Sparse Heterogeneous Mixtures," *The Annals of Statistics*, 32, 962–994. [594,598]

——— (2008), "Higher Criticism Thresholding: Optimal Feature Selection When Useful Features Are Rare and Weak," *Proceedings of the National Academy of Sciences of the United States of America*, 105, 14790–14795. [597]

——— (2015), "Special Invited Paper: Higher Criticism for Large-Scale Inference, Especially for Rare and Weak Effects," *Statistical Science*, 30, 1–25. [594,598,599]

Drusvyatskiy, D., Ioffe, A., and Lewis, A. (2015), "Transversality and Alternating Projections for Nonconvex Sets," *Foundations of Computational Mathematics*, 15, 1637–1651. [603]

Fama, E. F., and French, K. R. (1993), "Common Risk Factors in the Returns on Stocks and Bonds," *Journal of Financial Economics*, 33, 3–56. [592,593]

Fan, J., and Fan, Y. (2008), "High Dimensional Classification Using Features Annealed Independence Rules," *The Annals of Statistics*, 36, 2605. [593,597]

Fan, J., Han, X., and Gu, W. (2012), "Estimating False Discovery Proportion Under Arbitrary Covariance Dependence," *Journal of the American Statistical Association*, 107, 1019–1035. [592,594,598,600]

Fan, J., Liao, Y., and Liu, H. (2016), "An Overview of the Estimation of Large Covariance and Precision Matrices," *The Econometrics Journal*, 19, C1–C32. [593]

Fan, J., Liao, Y., and Mincheva, M. (2011), "High Dimensional Covariance Matrix Estimation in Approximate Factor Models," *The Annals of Statistics*, 39, 3320. [593]

——— (2013), "Large Covariance Estimation by Thresholding Principal Orthogonal Complements," *Journal of the Royal Statistical Society*, Series B, 75, 603–680. [593,594,605]

Fan, J., Xue, L., and Yao, J. (2017), "Sufficient Forecasting Using Factor Models," *Journal of Econometrics*, 201, 292–306. [606]

Fan, J., Xue, L., and Zou, H. (2014), "Strong Oracle Optimality of Folded Concave Penalized Estimation," *The Annals of Statistics*, 42, 819. [593]

Fan, Y., Jin, J., and Yao, Z. (2013), "Optimal Classification in Sparse Gaussian Graphic Model," *The Annals of Statistics*, 41, 2537–2571. [597]

Feingold, D. G., and Varga, R. S. (1962), "Block Diagonally Dominant Matrices and Generalizations of the Gerschgorin Circle Theorem," *Pacific Journal of Mathematics*, 12, 1241–1250. [605]

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse Inverse Covariance Estimation With the Graphical Lasso," *Biostatistics*, 9, 432–441. [594]

Gordon, G. J., Jensen, R. V., Hsiao, L.-L., Gullans, S. R., Blumenstock, J. E., Ramaswamy, S., Richards, W. G., Sugarbaker, D. J., and Bueno, R. (2002), "Translation of Microarray Data Into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelioma," *Cancer Research*, 62, 4963–4967. [597]

Hall, P., and Jin, J. (2008), "Properties of Higher Criticism Under Strong Dependence," *The Annals of Statistics*, 36, 381–402. [598,599]

——— (2010), "Innovated Higher Criticism for Detecting Sparse Signals in Correlated Noise," *The Annals of Statistics*, 38, 1686–1732. [594,598,599]

Horn, R. A., and Johnson, C. R. (2012), *Matrix Analysis*, Cambridge: Cambridge University Press. [593,596]

Huang, S., Jin, J., and Yao, Z. (2016), "Partial Correlation Screening for Estimating Large Precision Matrices, With Applications to Classification," *The Annals of Statistics*, 44, 2018–2057. [597]

Jager, L., and Wellner, J. A. (2007), "Goodness-of-Fit Tests via Phi-Divergences," *The Annals of Statistics*, 35, 2018–2053. [598]

Jin, J., and Wang, W. (2016), "Influential Features PCA for High Dimensional Clustering," *The Annals of Statistics*, 44, 2323–2359. [597]

Ke, Z. T., and Yang, F. (2017), "Covariate Assisted Variable Ranking," arXiv no. 1705.10370. [604]

Leek, J. T., and Storey, J. D. (2008), "A General Framework for Multiple Testing Dependence," *Proceedings of the National Academy of Sciences of the United States of America*, 105, 18718–18723. [592,598]

Lin, T., Ma, S., and Zhang, S. (2018), "Global Convergence of Unmodified 3-Block ADMM for a Class of Convex Minimization Problems," *Journal of Scientific Computing*, 76, 69–88. [602]

Ma, S., Xue, L., and Zou, H. (2013), "Alternating Direction Methods for Latent Variable Gaussian Graphical Model Selection," *Neural Computation*, 25, 2172–2198. [602]

Ma, Y., and Zhu, L. (2013), "A Review on Dimension Reduction," *International Statistical Review*, 81, 134–150. [606]

Mendoza, M., Raydan, M., and Tarazaga, P. (1998), "Computing the Nearest Diagonally Dominant Matrix," *Numerical Linear Algebra With Applications*, 5, 461–474. [593,601,602,605]

Netrapalli, P., Niranjan, U., Sanghavi, S., Anandkumar, A., and Jain, P. (2014), "Non-Convex Robust PCA," in *Advances in Neural Information Processing Systems*, pp. 1107–1115. [602]

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006), "Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies," *Nature Genetics*, 38, 904. [593]

Rothman, A. J., Levina, E., and Zhu, J. (2009), "Generalized Thresholding of Large Covariance Matrices," *Journal of the American Statistical Association*, 104, 177–186. [594]

Simes, R. J. (1986), "An Improved Bonferroni Procedure for Multiple Tests of Significance," *Biometrika*, 73, 751–754. [598]

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002), "Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression," *Proceedings of the National Academy of Sciences of the United States of America*, 99, 6567–6572. [597]

Tong, X., Feng, Y., and Zhao, A. (2016), "A Survey on Neyman-Pearson Classification and Suggestions for Future Research," *Wiley Interdisciplinary Reviews: Computational Statistics*, 8, 64–81. [606]

Wang, S., and Zhu, J. (2008), "Variable Selection for Model-Based High-Dimensional Clustering and Its Application to Microarray Data," *Biometrics*, 64, 440–448. [606]

Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M. E., and Yu, J. (2005), "Gene-Expression Profiles to Predict Distant Metastasis of Lymph-Node-Negative Primary Breast Cancer," *The Lancet*, 365, 671–679. [597]

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011), "Rare-Variant Association Testing for Sequencing Data With the Sequence Kernel Association Test," *The American Journal of Human Genetics*, 89, 82–93. [594,598]

Xue, L., Ma, S., and Zou, H. (2012), "Positive-Definite ℓ_1-Penalized Estimation of Large Covariance Matrices," *Journal of the American Statistical Association*, 107, 1480–1491. [594]

Zhu, J., and Hastie, T. (2005), "Kernel Logistic Regression and the Import Vector Machine," *Journal of Computational and Graphical Statistics*, 14, 185–205. [606]

Zou, H., and Li, R. (2008), "One-Step Sparse Estimates in Nonconcave Penalized Likelihood Models" (with discussion), *The Annals of Statistics*, 36, 1509. [593]

Zou, H., and Xue, L. (2018), "A Selective Overview of Sparse Principal Component Analysis," *Proceedings of the IEEE*, 106, 1311–1320. [606]