ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2018)
80, *Part* 5, *pp.* 1057–1086

Sparse generalized eigenvalue problem: optimal statistical rates via truncated Rayleigh flow

Kean Ming Tan,

University of Minnesota, Minneapolis, USA

Zhaoran Wang and Han Liu

Northwestern University, Evanston, USA

and Tong Zhang

Tencent Technology Shenzhen, People's Republic of China

[Received May 2016. Final revision July 2018]

Summary. The sparse generalized eigenvalue problem (GEP) plays a pivotal role in a large family of high dimensional statistical models, including sparse Fisher's discriminant analysis, canonical correlation analysis and sufficient dimension reduction. The sparse GEP involves solving a non-convex optimization problem. Most existing methods and theory in the context of specific statistical models that are special cases of the sparse GEP require restrictive structural assumptions on the input matrices. We propose a two-stage computational framework to solve the sparse GEP. At the first stage, we solve a convex relaxation of the sparse GEP. Taking the solution as an initial value, we then exploit a non-convex optimization perspective and propose the truncated Rayleigh flow method (which we call 'rifle') to estimate the leading generalized eigenvector. We show that rifle converges linearly to a solution with the optimal statistical rate of convergence. Theoretically, our method significantly improves on the existing literature by eliminating structural assumptions on the input matrices. To achieve this, our analysis involves two key ingredients: a new analysis of the gradient-based method on non-convex objective functions, and a fine-grained characterization of the evolution of sparsity patterns along the solution path. Thorough numerical studies are provided to validate the theoretical results.

Keywords: Convex relaxation; Non-convex optimization; Sparse canonical correlation analysis; Sparse Fisher's discriminant analysis; Sparse sufficient dimension reduction

1. Introduction

A large class of high dimensional statistical problems such as canonical correlation analysis (CCA), Fisher's discriminant analysis (FDA) and sufficient dimension reduction can be formulated as the generalized eigenvalue problem (GEP). Let $\mathbf{A} \in \mathbb{R}^{d\times d}$ be a symmetric matrix and let $\mathbf{B} \in \mathbb{R}^{d\times d}$ be a positive definite matrix. For a symmetric definite matrix pair $(\mathbf{A}, \mathbf{B})$, the GEP aims to obtain $\mathbf{v}^* \in \mathbb{R}^d$ satisfying

$$\mathbf{A}\mathbf{v}^* = \lambda_{\max}(\mathbf{A}, \mathbf{B}) \cdot \mathbf{B}\mathbf{v}^*, \tag{1}$$

where $\mathbf{v}^*$ is the leading generalized eigenvector corresponding to the largest generalized eigenvalue $\lambda_{\max}(\mathbf{A}, \mathbf{B})$ of the matrix pair $(\mathbf{A}, \mathbf{B})$. The largest generalized eigenvalue can also be characterized as

Address for correspondence: Kean Ming Tan, School of Statistics, University of Minnesota, 313 Ford Hall, 224 Church Street South East, Minneapolis, MN 55108, USA.
E-mail: ktan@umn.edu

1369–7412/18/801057

$$\lambda_{\max}(\mathbf{A},\mathbf{B}) = \max_{\mathbf{v}\in\mathbb{R}^d} \mathbf{v}^{\mathrm{T}}\mathbf{A}\mathbf{v}, \qquad \text{subject to } \mathbf{v}^{\mathrm{T}}\mathbf{B}\mathbf{v}=1.$$

In many real world applications, the matrix pair $(\mathbf{A},\mathbf{B})$ is a population quantity that is unknown in general. Instead, we can access only $(\hat{\mathbf{A}},\hat{\mathbf{B}})$, which is an estimator of $(\mathbf{A},\mathbf{B})$ based on n independent observations:

$$\hat{\mathbf{A}} = \mathbf{A} + \mathbf{E}_{\mathbf{A}}$$

and

$$\hat{\mathbf{B}} = \mathbf{B} + \mathbf{E}_{\mathbf{B}},$$

where $\mathbf{E}_{\mathbf{A}}$ and $\mathbf{E}_{\mathbf{B}}$ are stochastic errors due to finite sample estimation. For statistical models that are considered in this paper, $\mathbf{E}_{\mathbf{A}}$ and $\mathbf{E}_{\mathbf{B}}$ are symmetric matrices.

In the high dimensional setting in which $d>n$, we assume that the leading generalized eigenvector $\mathbf{v}^*$ is sparse. Let $s=\|\mathbf{v}^*\|_0$ be the number of non-zero entries in $\mathbf{v}^*$, and assume that s is much smaller than n and d. We aim to estimate a sparse $\mathbf{v}^*$ based on $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ by solving the optimization problem

$$\underset{\mathbf{v}\in\mathbb{R}^d}{\text{maximize}}\ \mathbf{v}^{\mathrm{T}}\hat{\mathbf{A}}\mathbf{v}, \qquad \text{subject to } \mathbf{v}^{\mathrm{T}}\hat{\mathbf{B}}\mathbf{v}=1, \quad \|\mathbf{v}\|_0 \leqslant s. \tag{2}$$

There are three major challenges in solving problem (2). Firstly, in the high dimensional setting, $\hat{\mathbf{B}}$ is singular and not invertible, and classical algorithms which require taking the inverse of $\hat{\mathbf{B}}$ are not directly applicable (Golub and Van Loan, 2012). Secondly, because of the normalization term $\mathbf{v}^{\mathrm{T}}\hat{\mathbf{B}}\mathbf{v}=1$, many recent proposals for solving sparse eigenvalue problems such as the truncated power method in Yuan and Zhang (2013) cannot be directly applied to solve problem (2). Thirdly, problem (2) requires maximizing a convex objective function over a non-convex set, which is 'NP' hard even when $\hat{\mathbf{B}}$ is the identity matrix (Moghaddam *et al.*, 2006a, b).

In this paper, we propose a two-stage computational framework for solving the sparse GEP (2). At the first stage, we solve a convex relaxation of problem (2). Our proposal generalizes the convex relaxation that was proposed in Gao *et al.* (2017) in the context of sparse CCA to the sparse GEP setting. Gao *et al.* (2017) assumed that $\mathbf{A}$ is low rank and positive semidefinite, and the rank of $\mathbf{A}$ is known. Our theoretical analysis removes all of those assumptions. Using the solution as an initial value, we propose a non-convex optimization algorithm to solve problem (2) directly. The algorithm proposed iteratively performs a gradient ascent step on the generalized Rayleigh quotient $\mathbf{v}^{\mathrm{T}}\hat{\mathbf{A}}\mathbf{v}/\mathbf{v}^{\mathrm{T}}\hat{\mathbf{B}}\mathbf{v}$, and a truncation step that preserves the top k entries of $\mathbf{v}$ with the largest magnitudes while setting the remaining entries to 0. Here, k is a tuning parameter that controls the cardinality of the solution. Theoretical guarantees are established for the proposed non-convex algorithm. To the best of our knowledge, this is the first general theoretical result for sparse GEPs in the high dimensional setting.

We provide a brief description of the theoretical result for the non-convex algorithm at the second stage. Let $\{\mathbf{v}_t\}_{t=0}^{L}$ be the solution sequence resulting from the algorithm proposed, where L is the total number of iterations and $\mathbf{v}_0$ is the initialization point. We prove that, under mild conditions,

$$\|\mathbf{v}^t - \mathbf{v}^*\|_2 \leqslant \underbrace{\nu^t\|\mathbf{v}^0 - \mathbf{v}^*\|_2}_{\text{optimization error}} + \underbrace{\frac{\surd\{\rho(\mathbf{E}_{\mathbf{A}},2k+s)^2 + \rho(\mathbf{E}_{\mathbf{B}},2k+s)^2\}}{\xi(\mathbf{A},\mathbf{B})}}_{\text{statistical error}} \qquad (t=1,\ldots,L). \tag{3}$$

The quantities $\nu\in(0,1)$ and $\xi(\mathbf{A},\mathbf{B})$ depend on the population matrix pair $(\mathbf{A},\mathbf{B})$. These quantities will be specified in Section 4. Meanwhile, $\rho(\mathbf{E}_{\mathbf{A}},2k+s)$ is defined as

$$\rho(\mathbf{E_A}, 2k+s) = \sup_{\|\mathbf{u}\|_2=1, \|\mathbf{u}\|_0 \leqslant 2k+s} |\mathbf{u}^{\mathrm{T}} \mathbf{E_A} \mathbf{u}| \tag{4}$$

and $\rho(\mathbf{E_B}, 2k+s)$ is defined similarly. The first term on the right-hand side quantifies the exponential decay of the optimization error, whereas the second term characterizes the statistical error due to finite sample estimation. In particular, for many statistical models that can be formulated as a sparse GEP such as sparse CCA, sparse FDA and sparse sufficient dimension reduction, it can be shown that

$$\max\{\rho(\mathbf{E_A}, 2k+s), \rho(\mathbf{E_B}, 2k+s)\} \leqslant \sqrt{\left\{\frac{(s+2k)\log(d)}{n}\right\}} \tag{5}$$

with high probability. Consequently, for any properly chosen k that is of the same order as s, the algorithm achieves an estimator of $\mathbf{v}^*$ with the optimal statistical rate of convergence $\sqrt{\{s\log(d)/n\}}$.

The sparse GEP (2) is also closely related to the classical matrix computation literature (see, for example, Golub and Van Loan (2012) for a survey, and more recent results in Ge *et al.* (2016)). There are two key differences between our results and existing work. Firstly, we have an additional non-convex constraint on the sparsity level, which allows us to handle the high dimensional setting. Secondly, because of the existence of stochastic errors, we allow the normalization matrix $\hat{\mathbf{B}}$ to be rank deficient, whereas in the classical setting $\hat{\mathbf{B}}$ is assumed to be positive definite. In comparison with existing generalized eigenvalue algorithms, our algorithm keeps the iterative solution sequence within a basin that involves only a few co-ordinates of $\mathbf{v}$ such that the corresponding submatrix of $\hat{\mathbf{B}}$ is positive definite. Moreover, our algorithm ensures that the statistical errors in result (3) are in terms of the largest sparse eigenvalues of the stochastic errors $\mathbf{E_A}$ and $\mathbf{E_B}$, which are defined in equation (4). In contrast, a straightforward application of classical matrix perturbation theory gives statistical error terms that involve the largest eigenvalues of $\mathbf{E_A}$ and $\mathbf{E_B}$, which are much larger than their sparse eigenvalues (Stewart and Sun, 1990).

An R package `rifle` for fitting and solving the sparse GEP can be found on the Comprehensive R Archive Network.

1.1. Notation

Let $\mathbf{v} = (v_1, \ldots, v_d)^{\mathrm{T}} \in \mathbb{R}^d$. We define the l_q-norm of $\mathbf{v}$ as $\|\mathbf{v}\|_q = (\Sigma_{j=1}^d |v_j|^q)^{1/q}$ for $1 \leqslant q < \infty$. Let $\lambda_{\max}(\mathbf{Z})$ and $\lambda_{\min}(\mathbf{Z})$ be the largest and smallest eigenvalues correspondingly. If $\mathbf{Z}$ is positive definite, we define its condition number as $\kappa(\mathbf{Z}) = \lambda_{\max}(\mathbf{Z})/\lambda_{\min}(\mathbf{Z})$. We denote $\lambda_k(\mathbf{Z})$ to be the kth eigenvalue of $\mathbf{Z}$, and the spectral norm of $\mathbf{Z}$ by $\|\mathbf{Z}\|_2 = \sup_{\|\mathbf{v}\|_2=1} \|\mathbf{Z}\mathbf{v}\|_2$. Furthermore, let $\|\mathbf{Z}\|_{1,1} = \Sigma_{i,j} |Z_{ij}|$, $\|\mathbf{Z}\|_{\infty,\infty} = \max_{i,j} |Z_{ij}|$ and $\|\mathbf{Z}\|_* = \mathrm{tr}(\mathbf{Z})$. For $F \subset \{1, \ldots, d\}$, let $\mathbf{Z}_{\cdot F} \in \mathbb{R}^{d\times|F|}$ and $\mathbf{Z}_{F\cdot} \in \mathbb{R}^{|F|\times d}$ be the submatrix of $\mathbf{Z}$ where the columns and rows are restricted to the set F. With some abuse of notation, let $\mathbf{Z}_F \in \mathbb{R}^{|F|\times|F|}$ be the submatrix of $\mathbf{Z}$, where the rows and columns are restricted to the set F. Finally, we define $\rho(\mathbf{Z}, s) = \sup_{\|\mathbf{u}\|_2=1, \|\mathbf{u}\|_0 \leqslant s} |\mathbf{u}^{\mathrm{T}} \mathbf{Z} \mathbf{u}|$.

2. Sparse generalized eigenvalue problem and its applications

Many high dimensional multivariate statistics methods can be formulated as special instances of problem (2). For instance, when $\hat{\mathbf{B}} = \mathbf{I}$, problem (2) reduces to the sparse principal component analysis (PCA) that has received considerable attention within the past decade (among others, Zou *et al.* (2006), d'Aspremont *et al.* (2007, 2008), Witten *et al.* (2009), Ma (2013), Cai *et al.* (2013), Yuan and Zhang (2013), Vu *et al.* (2013), Vu and Lei (2013), Birnbaum *et al.* (2013), Wang *et al.* (2013, 2014) and Gu *et al.* (2014)). In what follows, we provide three examples when $\hat{\mathbf{B}}$ is not the identity matrix. We start with sparse FDA for the classification problem (among others, Tibshirani *et al.* (2003), Guo *et al.* (2007), Leng (2008), Clemmensen *et al.* (2012), Mai *et al.* (2012, 2015), Kolar and Liu (2015), Gaynanova and Kolar (2015) and Fan *et al.* (2015)).

2.1. Example 1: sparse Fisher's discriminant analysis

Given n observations with K distinct classes, Fisher's discriminant problem seeks a low dimensional projection of the observations such that the between-class variance $\mathbf{\Sigma}_{\mathrm{b}}$ is large relative to the within-class variance $\mathbf{\Sigma}_{\mathrm{w}}$. Let $\hat{\mathbf{\Sigma}}_{\mathrm{b}}$ and $\hat{\mathbf{\Sigma}}_{\mathrm{w}}$ be estimators of $\mathbf{\Sigma}_{\mathrm{b}}$ and $\mathbf{\Sigma}_{\mathrm{w}}$ respectively. To obtain a sparse leading discriminant vector, we solve

$$\underset{\mathbf{v}}{\text{maximize}}\ \mathbf{v}^{\mathrm{T}}\hat{\mathbf{\Sigma}}_{\mathrm{b}}\mathbf{v}, \qquad \text{subject to } \mathbf{v}^{\mathrm{T}}\hat{\mathbf{\Sigma}}_{\mathrm{w}}\mathbf{v}=1, \quad \|\mathbf{v}\|_0 \leqslant s. \tag{6}$$

This is a special case of problem (2) with $\hat{\mathbf{A}}=\hat{\mathbf{\Sigma}}_{\mathrm{b}}$ and $\hat{\mathbf{B}}=\hat{\mathbf{\Sigma}}_{\mathrm{w}}$.

Next, we consider sparse CCA that explores the relationship between two high dimensional random vectors (among others Witten *et al.* (2009), Chen *et al.* (2013) and Gao *et al.* (2015, 2017)).

2.2. Example 2: sparse canonical correlation analysis

Let $\mathbf{X}$ and $\mathbf{Y}$ be two random vectors. Let $\mathbf{\Sigma}_x$ and $\mathbf{\Sigma}_y$ be the covariance matrices for $\mathbf{X}$ and $\mathbf{Y}$ respectively, and let $\mathbf{\Sigma}_{xy}$ be the cross-covariance matrix between $\mathbf{X}$ and $\mathbf{Y}$. To obtain sparse leading canonical direction vectors, we solve

$$\underset{\mathbf{v}_x,\mathbf{v}_y}{\text{maximize}}\ \mathbf{v}_x^{\mathrm{T}}\hat{\mathbf{\Sigma}}_{xy}\mathbf{v}_y, \qquad \text{subject to } \mathbf{v}_x^{\mathrm{T}}\hat{\mathbf{\Sigma}}_x\mathbf{v}_x=\mathbf{v}_y^{\mathrm{T}}\hat{\mathbf{\Sigma}}_y\mathbf{v}_y=1, \quad \|\mathbf{v}_x\|_0\leqslant s_x, \quad \|\mathbf{v}_y\|_0\leqslant s_y, \tag{7}$$

where s_x and s_y control the cardinality of $\mathbf{v}_x$ and $\mathbf{v}_y$. This is a special case of problem (2) with

$$\hat{\mathbf{A}}=\begin{pmatrix}\mathbf{0} & \hat{\mathbf{\Sigma}}_{xy}\\ \hat{\mathbf{\Sigma}}_{xy} & \mathbf{0}\end{pmatrix},$$
$$\hat{\mathbf{B}}=\begin{pmatrix}\hat{\mathbf{\Sigma}}_x & \mathbf{0}\\ \mathbf{0} & \hat{\mathbf{\Sigma}}_y\end{pmatrix},$$
$$\mathbf{v}=\begin{pmatrix}\mathbf{v}_x\\ \mathbf{v}_y\end{pmatrix}.$$

Theoretical guarantees for sparse CCA have been established recently. Chen *et al.* (2013) proposed a non-convex optimization algorithm for solving problem (7) with theoretical guarantees. However, their algorithm involves obtaining accurate estimators of $\mathbf{\Sigma}_x^{-1}$ and $\mathbf{\Sigma}_y^{-1}$, which is in general difficult to do without imposing sparsity assumptions on $\mathbf{\Sigma}_x^{-1}$ and $\mathbf{\Sigma}_y^{-1}$. In a follow-up work, Gao *et al.* (2017) proposed a two-stage procedure that attains the optimal statistical rate of convergence (Gao *et al.*, 2015). However, they required the matrix $\mathbf{\Sigma}_{xy}$ to be low rank and positive semidefinite, and that the rank of $\mathbf{\Sigma}_{xy}$ is known *a priori*. As suggested in Gao *et al.* (2015), the low rank assumption on $\mathbf{\Sigma}_{xy}$ may be unrealistic in many real data applications where one is interested in recovering the first few sparse canonical correlation directions whereas there might be additional directions in the population structure. Our proposal does not impose any structural assumptions on $\mathbf{\Sigma}_x$ and $\mathbf{\Sigma}_y$, and we only require $\mathbf{\Sigma}_{xy}$ to be approximately low rank in the sense that the leading generalized eigenvalue is larger than the remaining generalized eigenvalues.

Next, we consider a regression problem with a univariate response Y and d-dimensional covariates $\mathbf{X}$, with the goal of inferring the conditional distribution of Y given $\mathbf{X}$. Sufficient dimension reduction is a popular approach for reducing the dimensionality of the covariates (Li, 1991; Cook and Lee, 1999; Cook, 2000, 2007; Cook and Forzani, 2008; Ma and Zhu, 2013). It can be shown that many sufficient dimension reduction methods can be formulated as GEPs (Li, 2007; Chen *et al.*, 2010). In what follows we consider sparse sliced inverse regression (Li, 1991).

2.3. Example 3: sparse sliced inverse regression

Consider the model

$$Y = f(\mathbf{v}_1^{\mathrm{T}}\mathbf{X}, \ldots, \mathbf{v}_K^{\mathrm{T}}\mathbf{X}, \epsilon),$$

where ϵ is the stochastic error independent of $\mathbf{X}$, and $f(\cdot)$ is an unknown link function. Li (1991) proved that, under regularity conditions, the subspace that is spanned by $\mathbf{v}_1, \ldots, \mathbf{v}_K$ can be identified. Let Σ_x be the covariance matrix for $\mathbf{X}$ and let $\Sigma_{E(\mathbf{X}|Y)}$ be the covariance matrix of the conditional expectation $E(\mathbf{X}|Y)$. The first leading eigenvector of the subspace that is spanned by $\mathbf{v}_1, \ldots, \mathbf{v}_K$ can be identified by solving

$$\text{maximize}_{\mathbf{v}}\ \mathbf{v}^{\mathrm{T}}\hat{\Sigma}_{E(\mathbf{X}|Y)}\mathbf{v}, \qquad \text{subject to } \mathbf{v}^{\mathrm{T}}\hat{\Sigma}_x\mathbf{v} = 1, \quad \|\mathbf{v}\|_0 \leqslant s. \tag{8}$$

This is a special case of problem (2) with $\hat{\mathbf{A}} = \hat{\Sigma}_{E(\mathbf{X}|Y)}$ and $\hat{\mathbf{B}} = \hat{\Sigma}_x$.

Many researchers have proposed methods for sparse sliced inverse regression (Li and Nachtsheim, 2006; Zhu *et al.*, 2006; Li and Yin, 2008; Chen *et al.*, 2010; Yin and Hilafu, 2015). More generally, in the context of sparse sufficient dimension reduction, Li (2007) and Chen *et al.* (2010) reformulated sparse sufficient dimension reduction problems into the sparse GEP (2). However, these approaches lack algorithmic and non-asymptotic statistical guarantees in the high dimensional setting. Our results are applicable to most sparse sufficient dimension reduction methods.

3. Methodology and algorithm

In Section 3.1, we propose an iterative algorithm to estimate $\mathbf{v}^*$ by solving problem (2), which we refer to as the truncated Rayleigh flow method (and call rifle). Rifle requires an input of an initial vector $\mathbf{v}_0$ that is sufficiently close to $\mathbf{v}^*$. For this, we propose a convex optimization approach to obtain such an initial vector $\mathbf{v}_0$ in Section 3.2.

3.1. Truncated Rayleigh flow method rifle

Optimization problem (2) can be rewritten as

$$\underset{\mathbf{v}\in\mathbb{R}^d}{\text{maximize}}\ \frac{\mathbf{v}^{\mathrm{T}}\hat{\mathbf{A}}\mathbf{v}}{\mathbf{v}^{\mathrm{T}}\hat{\mathbf{B}}\mathbf{v}}, \qquad \text{subject to } \|\mathbf{v}\|_0 \leqslant s,$$

where the objective function is referred to as the generalized Rayleigh quotient.

The crux of our proposed algorithm is as follows. Given an initial vector $\mathbf{v}_0$, we first compute the gradient of the generalized Rayleigh quotient. We then update the initial vector by its ascent direction and normalize it such that the updated vector has norm 1. This step ensures that the generalized Rayleigh quotient for the updated vector is at least as large as that of the initial

Table 1. Algorithm 1: truncated Rayleigh flow method (rifle)

```
Input: matrices $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$, initial vector $\mathbf{v}_0$, cardinality
  $k \in \{1, \ldots, d\}$ and step size $\eta$
Truncate: truncate $\mathbf{v}_0$ by keeping the largest $k$ absolute
  elements, and setting the remaining entries to 0
Let $t = 1$ and repeat the following steps until convergence:
  1, $\rho_{t-1} \leftarrow \mathbf{v}_{t-1}^{\mathrm{T}} \hat{\mathbf{A}} \mathbf{v}_{t-1} / \mathbf{v}_{t-1}^{\mathrm{T}} \hat{\mathbf{B}} \mathbf{v}_{t-1}$;
  2, $\mathbf{C} \leftarrow \mathbf{I} + (\eta/\rho_{t-1})(\hat{\mathbf{A}} - \rho_{t-1}\hat{\mathbf{B}})$;
  3, $\mathbf{v}'_t \leftarrow \mathbf{C}\mathbf{v}_{t-1} / \|\mathbf{C}\mathbf{v}_{t-1}\|_2$;
  4, let $F_t = \mathrm{supp}(\mathbf{v}'_t, k)$ contain the indices of $\mathbf{v}'_t$ with the largest
     $k$ absolute values and truncate($\mathbf{v}'_t$, $F_t$) be the truncated
     vector of $\mathbf{v}'_t$ by setting $(\mathbf{v}'_t)_i = 0$ for $i \notin F_t$;
  5, $\hat{\mathbf{v}}_t \leftarrow \mathrm{truncate}(\mathbf{v}'_t, F_t)$;
  6, $\mathbf{v}_t \leftarrow \hat{\mathbf{v}}_t / \|\hat{\mathbf{v}}_t\|_2$;
  7, $t \leftarrow t + 1$
Output: $\mathbf{v}_t$
```

vector. Indeed, in theorem 1, we show that, if the initial vector $\mathbf{v}_0$ is close to $\mathbf{v}^*$, then this step ensures that the updated vector is closer to $\mathbf{v}^*$ compared with $\mathbf{v}_0$. Next, we truncate the updated vector by keeping the elements with the largest k absolute values and setting the remaining elements to 0. This step ensures that the updated vector is k sparse, i.e. only k entries are non-zero. Finally, we normalize the updated vector such that it has norm 1. These steps are repeated until convergence. We summarize the details in algorithm 1 (Table 1).

In addition to an initial vector $\mathbf{v}_0$, algorithm 1 requires the choice of a step size η and a tuning parameter k on the cardinality of the solution. As suggested by the theoretical results in Section 4, we need η to be sufficiently small such that $\eta \lambda_{\max}(\hat{\mathbf{B}}) < 1$. In practice, the tuning parameter k can be selected by using cross-validation or based on prior knowledge. The computational complexity for each iteration of algorithm 1 is $\mathcal{O}(kd + d)$: $\mathcal{O}(d)$ for selecting the k largest elements of a d-dimensional vector to obtain the set F_t, and $\mathcal{O}(kd)$ for taking the product between a truncated vector and a matrix with columns restricted to the set F_t, and for calculating the difference between two matrices with columns restricted to the set F_t.

3.2. A convex optimization approach to obtain $\mathbf{v}_0$

As mentioned in Section 3.1, it is crucial to obtain an initial vector $\mathbf{v}_0$ that is close to $\mathbf{v}^*$ for rifle. Gao *et al.* (2017) have proposed a convex formulation to estimate the subspace that is spanned by the K leading generalized eigenvectors for sparse CCA, under the assumption that $\mathbf{A}$ is low rank and positive semidefinite. Rather than estimating the K leading generalized eigenvectors, the main idea of Gao *et al.* (2017) is to obtain an estimator of the subspace spanned by the K leading generalized eigenvectors directly. In this section, we point out that the convex relaxation proposed can be used more generally to estimate the subspace of a sparse generalized eigenvalue problem, without the low rank and positive semidefinite structural assumptions on $\mathbf{A}$.

Similarly to problem (2), the optimization problem for estimating the K generalized eigenvectors can be written as

$$\underset{\mathbf{U} \in \mathbb{R}^{d \times K}}{\text{minimize}} \; -\mathrm{tr}(\mathbf{U}^{\mathrm{T}} \hat{\mathbf{A}} \mathbf{U}), \qquad \text{subject to } \mathbf{U}^{\mathrm{T}} \hat{\mathbf{B}} \mathbf{U} = \mathbf{I}_K.$$

Rather than estimating the K generalized eigenvectors which involves minimizing a concave function, we consider approximating the subspace that is spanned by $\mathbf{U}$. Let $\mathbf{P} = \mathbf{U}\mathbf{U}^{\mathrm{T}}$ and let

Table 2. Algorithm 2: alternating direction methods of multiplier algorithm for solving problem (10)

```
Input: matrices $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$, tuning parameters $\zeta$ and $K$, alternating
    direction methods of multiplier parameter $\nu$ and convergence
    criterion $\epsilon$
Initialize: matrices $\mathbf{P}_0$, $\mathbf{H}_0$ and $\boldsymbol{\Gamma}_0$
Let $t=1$ and repeat the following steps until $\|\mathbf{P}_{t+1}-\mathbf{P}_t\|_{\mathrm{F}} \leqslant \epsilon$:
    1, update $\mathbf{P}$ by solving the lasso problem
        $\mathbf{P}_{t+1}=\arg\min_{\mathbf{P}} \frac{\nu}{2}\|\hat{\mathbf{B}}^{1/2}\mathbf{P}\hat{\mathbf{B}}^{1/2}-\mathbf{H}_t+\boldsymbol{\Gamma}_t\|_{\mathrm{F}}^2-\mathrm{tr}(\hat{\mathbf{A}}\mathbf{P})+\zeta\|\mathbf{P}\|_{1,1}$;
    2, let $\Sigma_{j=1}^{d}\omega_j\mathbf{a}_j\mathbf{b}_j^{\mathrm{T}}$ be the singular value decomposition of
        $\boldsymbol{\Gamma}_t+\mathbf{B}^{1/2}\mathbf{P}_{t+1}\hat{\mathbf{B}}^{1/2}$ and let
        $\gamma^*=\arg\min_{\gamma>0}\gamma$    subject to $\sum_{j=1}^{d}\min\{1,\max(\omega_j-\gamma,0)\}\leqslant K$;
        update $\mathbf{H}$ by
            $\mathbf{H}_{t+1}=\sum_{j=1}^{d}\min\{1,\max(\omega_j-\gamma^*,0)\}\mathbf{a}_j\mathbf{b}_j^{\mathrm{T}}$;
    3, update $\boldsymbol{\Gamma}$ by
            $\boldsymbol{\Gamma}_{t+1}=\boldsymbol{\Gamma}_t+\hat{\mathbf{B}}^{1/2}\mathbf{P}_{t+1}\hat{\mathbf{B}}^{1/2}-\mathbf{H}_{t+1}$;
    4, $t\leftarrow t+1$
```

$\mathcal{O}=\{\hat{\mathbf{B}}^{1/2}\mathbf{P}\hat{\mathbf{B}}^{1/2}:\mathbf{U}^{\mathrm{T}}\hat{\mathbf{B}}\mathbf{U}=\mathbf{I}_K\}$. By a change of variable, the above problem can be rewritten as

$$\underset{\mathbf{P}\in\mathbb{R}^{d\times d}}{\text{minimize}} \; -\mathrm{tr}(\hat{\mathbf{A}}\mathbf{P}), \qquad \text{subject to } \mathbf{P}\in\mathcal{O}, \tag{9}$$

where the objective function is now linear in $\mathbf{P}$.

We consider the following convex relaxation of problem (9) with a lasso penalty to encourage the estimated subspace to be sparse:

$$\underset{\mathbf{P}}{\text{minimize}} \; -\mathrm{tr}(\hat{\mathbf{A}}\mathbf{P})+\zeta\|\mathbf{P}\|_{1,1}, \qquad \text{subject to}\|\hat{\mathbf{B}}^{1/2}\mathbf{P}\hat{\mathbf{B}}^{1/2}\|_* \leqslant K \text{ and } \|\hat{\mathbf{B}}^{1/2}\mathbf{P}\hat{\mathbf{B}}^{1/2}\|_2\leqslant 1, \tag{10}$$

where $\|\cdot\|_*$ and $\|\cdot\|_2$ are the nuclear norm and spectral norm that encourage the solution to be low rank and that its eigenvalue to be bounded respectively. Here, ζ and K are two tuning parameters that encourage the estimated subspace $\mathbf{P}$ to be sparse and low rank respectively. The convex optimization problem (10) can be solved by using the alternating direction methods of multiplier algorithm; we summarize the details in algorithm 2 (Table 2) (Boyd *et al.*, 2010; Eckstein, 2012). The computational bottleneck in algorithm 2 is the singular value decomposition on a $d\times d$ matrix, thus yielding a computational complexity of $\mathcal{O}(d^3)$. Compared with the computational complexity of $\mathcal{O}(kd+d)$ for algorithm 1, it can be seen that obtaining a good initial vector $\mathbf{v}_0$ is much more time consuming than refining the initial value.

Let $\hat{\mathbf{P}}$ be an estimator obtained from solving problem (10). Then, the initial value $\mathbf{v}_0$ can be set to be the largest eigenvector of $\hat{\mathbf{P}}$. The theoretical guarantees for $\mathbf{v}_0$ that are obtained via this approach are presented in proposition 1 in Section 4.1. In practice, for obtaining an initial value $\mathbf{v}_0$, we can simply set $K=1$ and ζ to be approximately $\surd\{\log(d)/n\}$. In fact, we suggest setting ζ conservatively since there is a refinement step using rifle to obtain an estimator that is closer to $\mathbf{v}^*$.

4. Theoretical results

We show that, if the matrix pair $(\mathbf{A}, \mathbf{B})$ has a unique sparse leading generalized eigenvector, then algorithm 1 can accurately recover the population leading generalized eigenvector from the noisy matrix pair $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$. Recall from Section 1 that $\mathbf{A}$ is symmetric and $\mathbf{B}$ is positive definite. This condition ensures that all generalized eigenvalues are real. Recall that $\mathbf{v}^*$ is the leading generalized eigenvector of $(\mathbf{A}, \mathbf{B})$. Let $V = \text{supp}(\mathbf{v}^*)$ be the index set corresponding to the non-zero elements of $\mathbf{v}^*$, and let $|V| = s$. Let $F \subset \{1, \ldots, d\}$ be a superset of V, i.e. $V \subset F$, with cardinality $|F| = k'$. Throughout the paper, for notational convenience, let λ_j and $\hat{\lambda}_j$ be the jth generalized eigenvalue of the matrix pairs $(\mathbf{A}, \mathbf{B})$ and $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ respectively. Moreover, let $\lambda_j(F)$ and $\hat{\lambda}_j(F)$ be the jth generalized eigenvalue of the matrix pairs $(\mathbf{A}_F, \mathbf{B}_F)$ and $(\hat{\mathbf{A}}_F, \hat{\mathbf{B}}_F)$ respectively.

Our theoretical results depend on several quantities that are specific to the generalized eigenvalue problem. Let

$$\text{cr}(\mathbf{A}, \mathbf{B}) = \min_{\mathbf{v}:\|\mathbf{v}\|_2=1} \{(\mathbf{v}^{\mathrm{T}}\mathbf{A}\mathbf{v})^2 + (\mathbf{v}^{\mathrm{T}}\mathbf{B}\mathbf{v})^2\}^{1/2} > 0 \tag{11}$$

be the Crawford number of the symmetric definite matrix pair $(\mathbf{A}, \mathbf{B})$ (Stewart, 1979). Let

$$\begin{aligned} \text{cr}(k') &= \inf_{F:|F| \leqslant k'} \text{cr}(\mathbf{A}_F, \mathbf{B}_F), \\ \epsilon(k') &= \surd\{\rho(\mathbf{E}_{\mathbf{A}}, k')^2 + \rho(\mathbf{E}_{\mathbf{B}}, k')^2\}, \end{aligned} \tag{12}$$

where $\rho(\mathbf{E}_{\mathbf{A}}, k')$ is as defined in equation (4). In what follows, we start with an assumption that these quantities are upper bounded for sufficiently large n.

Assumption 1. For sufficiently large n, there are constants $b, c > 0$ such that

$$\frac{\epsilon(k')}{\text{cr}(k')} \leqslant b$$

and

$$\rho(\mathbf{E}_{\mathbf{B}}, k') \leqslant c\,\lambda_{\min}(\mathbf{B})$$

for any $k' \ll n$, where $\text{cr}(k')$ and $\epsilon(k')$ are defined in expression (12).

Provided that n is sufficiently large, it can be shown that assumption 1 holds with high probability for most statistical models. In fact, we shall show in proposition 2 in Section 4.2 that, as long as $n > Ck' \log(d)$ for some sufficiently large constant C, then assumption 1 is satisfied with high probability for most statistical models. We shall use the following implications of assumption 1 in our theoretical analysis, which are implied by matrix perturbation theory (Stewart, 1979; Stewart and Sun, 1990). In detail, by applications of lemmas 1 and 2 in Appendix A, we have that, for any $F \subset \{1, \ldots, d\}$ with $|F| = k'$, there are constants a and c such that

$$\begin{gathered} (1-a)\lambda_j(F) \leqslant \hat{\lambda}_j(F) \leqslant\ (1+a)\lambda_j(F), \\ (1-c)\lambda_j(\mathbf{B}_F) \leqslant \lambda_j(\hat{\mathbf{B}}_F) \leqslant\ (1+c)\lambda_j(\mathbf{B}_F) \end{gathered}$$

and

$$c_{\text{lower}}\,\kappa(\mathbf{B}) \leqslant \kappa(\hat{\mathbf{B}}_F) \leqslant c_{\text{upper}}\,\kappa(\mathbf{B}), \tag{13}$$

where $c_{\text{lower}} = (1-c)/(1+c)$, $c_{\text{upper}} = (1+c)/(1-c)$, c is the same constant as in assumption 1 and $\kappa(\mathbf{B})$ is the condition number of the matrix $\mathbf{B}$. Meanwhile, let $\gamma = (1+a)\lambda_2/\{(1-a)\lambda_1\}$.

Finally, we define $\mathbf{v}(F)$ to be the solution of a GEP restricted to a superset of V ($V \subset F$):

$$\mathbf{v}(F) = \arg\max_{\mathbf{v}\in\mathbb{R}^d} \mathbf{v}^{\mathrm{T}}\hat{\mathbf{A}}\mathbf{v}, \qquad \text{subject to } \mathbf{v}^{\mathrm{T}}\hat{\mathbf{B}}\mathbf{v} = 1, \quad \operatorname{supp}(\mathbf{v}) \subseteq F. \tag{14}$$

The quantity $\mathbf{v}(F)$ can be interpreted as the solution of a GEP for a low dimensional problem when $k' < n$. In the following theorem, we present our main theoretical result for algorithm 1 as a function of the l_2-distance between $\mathbf{v}(F)$ and $\mathbf{v}^*$.

Theorem 1. Let $k' = 2k + s$ and choose $k = Cs$ for sufficiently large C. In addition, choose η such that $\eta\,\lambda_{\max}(\mathbf{B}) < 1/(1+c)$ and

$$\nu = \surd[1 + 2\{(s/k)^{1/2} + s/k\}] \sqrt{\left\{1 - \frac{1+c}{8}\eta\,\lambda_{\min}(\mathbf{B})\frac{1-\gamma}{c_{\text{upper}}\kappa(\mathbf{B}) + \gamma}\right\}} < 1.$$

Input an initial vector $\mathbf{v}_0$ with $\|\mathbf{v}_0\|_2 = 1$ satisfying $|(\mathbf{v}^*)^{\mathrm{T}}\mathbf{v}_0|/\|\mathbf{v}^*\|_2 \geqslant 1 - \theta(\mathbf{A},\mathbf{B})$, where $\theta(\mathbf{A},\mathbf{B})$ is a quantity given in lemma 3 that depends on the matrix pair $(\mathbf{A},\mathbf{B})$. Under assumption 1, we have

$$\sqrt{\left\{1 - \frac{|(\mathbf{v}^*)^{\mathrm{T}}\mathbf{v}_t|}{\|\mathbf{v}^*\|_2}\right\}} \leqslant \nu^t \surd\theta(\mathbf{A},\mathbf{B}) + \frac{\surd 20}{1-\nu}\sqrt{\left\{1 - \frac{|\mathbf{v}(F)^{\mathrm{T}}\mathbf{v}^*|}{\|\mathbf{v}(F)\|_2\|\mathbf{v}^*\|_2}\right\}}. \tag{15}$$

For simplicity, assume that $(\mathbf{v}^*)^{\mathrm{T}}\mathbf{v}_t$ is positive without loss of generality. Since $\mathbf{v}_t$ is a unit vector, from inequality (15) we have

$$1 - \frac{|(\mathbf{v}^*)^{\mathrm{T}}\mathbf{v}_t|}{\|\mathbf{v}^*\|_2} = \frac{1}{2}\left\|\mathbf{v}_t - \frac{\mathbf{v}^*}{\|\mathbf{v}^*\|_2}\right\|_2^2,$$
$$1 - \frac{|\mathbf{v}(F)^{\mathrm{T}}\mathbf{v}^*|}{\|\mathbf{v}(F)\|_2\|\mathbf{v}^*\|_2} = \frac{1}{2}\left\|\frac{\mathbf{v}(F)}{\|\mathbf{v}(F)\|_2} - \frac{\mathbf{v}^*}{\|\mathbf{v}^*\|_2}\right\|_2^2.$$

Thus, result (15) states that the l_2-distance between $\mathbf{v}^*/\|\mathbf{v}^*\|_2$ and $\mathbf{v}_t$ can be upper bounded by two terms. The first term on the right-hand side of inequality (15) quantifies the optimization error, which decreases to 0 at a geometric rate since $\nu < 1$. Meanwhile, the second term on the right-hand side of inequality (15) is the statistical error that is introduced for solving GEPs restricted to the set F as in problem (14). The result in theorem 1 depends on the estimation error between $\mathbf{v}(F)$ and $\mathbf{v}^*$. The following corollary quantifies such estimation error for a general class of symmetric definite matrix pairs $(\mathbf{A},\mathbf{B})$.

Corollary 1. For a general class of symmetric definite matrix pairs $(\mathbf{A},\mathbf{B})$, let

$$\Delta\lambda = \min_{j>1} \frac{\lambda_1 - (1+a)\lambda_j}{\surd(1+\lambda_1^2)\surd\{1 + (1-a)^2\lambda_j^2\}} \tag{16}$$

denote the eigengap for the GEP (Stewart, 1979; Stewart and Sun, 1990). Assume that $\Delta\lambda > \epsilon(k')/\mathrm{cr}(k')$. Then, under the same conditions as in theorem 1, we have

$$\sqrt{\left\{1 - \frac{|(\mathbf{v}^*)^{\mathrm{T}}\mathbf{v}_t|}{\|\mathbf{v}^*\|_2}\right\}} \leqslant \nu^t \surd\theta(\mathbf{A},\mathbf{B}) + \frac{\surd 10}{1-\nu}\frac{2}{\Delta\lambda\{\mathrm{cr}(k') - \epsilon(k')\}}\epsilon(k'),$$

where $\epsilon(k') = \surd\{\rho(\mathbf{E_A},k')^2 + \rho(\mathbf{E_B},k')^2\}$.

For a large class of statistical models, $\epsilon(k')$ converges to 0 at the rate of $\surd\{s\log(d)/n\}$ with high probability.

4.1. Theoretical justification for the initialization in problem (10)

Theorem 1 involves a condition on the initialization $\mathbf{v}_0$: the cosine of the angle between $\mathbf{v}^*$ and $\mathbf{v}_0$ needs to be strictly larger than a constant. In other words, the initialization $\mathbf{v}_0$ needs to be close to $\mathbf{v}^*$. We now present some theoretical guarantees for the initialization procedure in Section 3.2. In the context of sparse CCA, Gao *et al.* (2017) have shown that the estimated subspace that is obtained from solving a convex relaxation of the form (10) converges to the true subspace, under the assumption that $\mathbf{A}$ is low rank and positive semidefinite, and that the rank of $\mathbf{A}$ is known. In the following proposition, we remove those assumptions on $\mathbf{A}$. Thus, a similar result holds more generally for the sparse generalized eigenvalue problem with symmetric definite matrix pair $(\mathbf{A},\mathbf{B})$.

For this, we define some additional notation. Let $\mathbf{V}^* \in \mathbb{R}^{d\times d}$ be d generalized eigenvectors and let $\boldsymbol{\Lambda}^* \in \mathbb{R}^{d\times d}$ be a diagonal matrix of generalized eigenvalues of the matrix pair $(\mathbf{A},\mathbf{B})$. Let $\mathcal{S}_v$ be a set containing indices of non-zero rows of $\mathbf{V}^* \in \mathbb{R}^{d\times d}$. For simplicity, assume that $|\mathcal{S}_v| = s$ and that the eigenvalues of $\mathbf{B}$ are bounded. The matrix $\mathbf{A}$ can be rewritten in terms of its generalized eigenvectors and generalized eigenvalues up to sign jointly, $\mathbf{A} = \mathbf{B}\mathbf{V}^*\boldsymbol{\Lambda}^*(\mathbf{V}^*)^{\mathrm{T}}\mathbf{B}$ (Gao *et al.*, 2017). Let $\tilde{\mathbf{A}} = \hat{\mathbf{B}}\mathbf{V}^*\boldsymbol{\Lambda}^*(\mathbf{V}^*)^{\mathrm{T}}\hat{\mathbf{B}}$ and let $\mathbf{P}^* = \mathbf{V}^*_{\cdot K}(\mathbf{V}^*_{\cdot K})^{\mathrm{T}}$, where $\mathbf{V}^*_{\cdot K}$ are the first K generalized eigenvectors of $(\mathbf{A},\mathbf{B})$. Let $\hat{\mathbf{P}}$ be a solution to problem (10) with tuning parameters ζ and K. The following proposition establishes an upper bound for the difference between $\hat{\mathbf{P}}$ and $\mathbf{P}^*$ under the Frobenius norm.

Proposition 1. Assume that n is sufficiently large such that $\rho(\mathbf{E}_{\mathbf{B}}, s^2) \leqslant c\lambda_{\min}(\mathbf{B})$, where c is the same constant as appears in assumption 1. Let $\delta_{\text{gap}} = \lambda_K - c\kappa(\mathbf{B})\lambda_{K+1}/(1-c)$, and assume that $\delta_{\text{gap}} > 0$. Set $\zeta > 2\|\hat{\mathbf{A}} - \tilde{\mathbf{A}}\|_{\infty,\infty}$. Then,

$$\|\hat{\mathbf{P}} - \mathbf{P}^*\|_{\mathrm{F}} \leqslant C\left(\frac{s}{\delta_{\text{gap}}}\|\hat{\mathbf{A}} - \tilde{\mathbf{A}}\|_{\infty,\infty} + K\|\hat{\mathbf{B}}_{\mathcal{S}_v} - \mathbf{B}_{\mathcal{S}_v}\|_2\right),$$

where C is a generic constant that does not depend on the generalized eigenvalues and the dimensions n, d, s and K.

For most statistical models, it can be shown that $\|\hat{\mathbf{A}} - \tilde{\mathbf{A}}\|_{\infty,\infty} \leqslant C_1\surd\{\log(d)/n\}$ and $\|\hat{\mathbf{B}}_{\mathcal{S}_v} - \mathbf{B}_{\mathcal{S}_v}\|_2 \leqslant C_2\surd(s/n)$ with high probability for generic constants C_1 and C_2. Thus, picking $\zeta > C_3\surd\{\log(d)/n\}$, the upper bound can be simplified to

$$\|\hat{\mathbf{P}} - \mathbf{P}^*\|_{\mathrm{F}} \leqslant C\left[\frac{s}{\delta_{\text{gap}}}\sqrt{\left\{\frac{\log(d)}{n}\right\}} + K\sqrt{\left(\frac{s}{n}\right)}\right].$$

Choosing $K = 1$ in expression (10), by a variant of the Davis–Kahan theorem in Vu *et al.* (2013), proposition 1 guarantees that, by setting $\mathbf{v}_0$ to be the leading eigenvector of $\hat{\mathbf{P}}$, then $\mathbf{v}_0$ will be sufficiently close to $\mathbf{v}^*$ as long as the conditions in proposition 1 are satisfied. In the next section, we shall quantify the sample size condition that is needed for proposition 1 to hold under various statistical models.

4.2. Applications to sparse principal components analysis and sparse canonical correlation analysis

In this section, we provide some discussions on the implications of theorem 1 and proposition 1 in the context of sparse PCA and CCA. More specifically, we first verify that the initial vector $\mathbf{v}_0$ that is obtained from solving problem (10) is close to $\mathbf{v}^*$. Therefore, the assumption on $\mathbf{v}_0$ in theorem 1 is satisfied. Next, we compare our results from theorem 1 with the minimax optimal rate of convergence for the two statistical models.

4.2.1. *Sparse principal component analysis*

We start with the sparse PCA problem. We assume the model $\mathbf{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. As mentioned in Section 2, sparse PCA is a special case of the sparse GEP when $(\mathbf{A}, \mathbf{B}) = (\boldsymbol{\Sigma}, \mathbf{I})$ and $(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = (\hat{\boldsymbol{\Sigma}}, \mathbf{I})$, where $\hat{\boldsymbol{\Sigma}}$ is the sample covariance matrix. Thus, optimization problem (10) reduces to a convex relaxation of sparse PCA proposed by Vu *et al.* (2013). In this case, using a variant of the theoretical results in proposition 1, the initial value $\mathbf{v}_0$ converges to $\mathbf{v}^*$ as long as $n > Cs^2 \log(d)$. Applying corollary 1 directly to the sparse PCA problem will give a loose upper bound (on the eigenfactor) since the additional information on the matrix pair $(\mathbf{A}, \mathbf{B}) = (\boldsymbol{\Sigma}, \mathbf{I})$, with $\mathbf{B}$ restricted to the identity matrix and $\mathbf{A}$ restricted to positive definite matrices, are not used in the derivation of corollary 1. In other words, the results in corollary 1 are derived under a much larger class of matrix pair $(\mathbf{A}, \mathbf{B})$. For this, we resort to the following corollary on the variant of the Davis–Kahan perturbation result for sparse PCA (see, for instance, Yu *et al.* (2014)).

Corollary 2. Let $(\mathbf{A}, \mathbf{B}) = (\boldsymbol{\Sigma}, \mathbf{I})$ and let $\boldsymbol{\Sigma}$ be a symmetric positive definite matrix. Let $\hat{\mathbf{A}} = \hat{\boldsymbol{\Sigma}}$ be the sample covariance matrix. We have that

$$\rho(\hat{\mathbf{A}} - \mathbf{A}, s) \leqslant C\surd\lambda_1(\mathbf{A})\sqrt{\left\{\frac{s\log(d)}{n}\right\}}$$

holds with high probability for some constant $C > 0$. Suppose that $|F| = k'$ and that $k' = \mathcal{O}(s)$. Then, by the Davis–Kahan theorem,

$$\sqrt{\left\{1 - \frac{|\mathbf{v}(F)^{\mathrm{T}}\mathbf{v}^*|}{\|\mathbf{v}(F)\|_2\|\mathbf{v}^*\|_2}\right\}} \leqslant C' \frac{\surd\lambda_1(\mathbf{A})}{\lambda_1(\mathbf{A}) - \lambda_2(\mathbf{A})}\sqrt{\left\{\frac{s\log(d)}{n}\right\}}$$

holds with high probability for some constant $C' > 0$.

Combining corollary 2 with theorem 1, our results indicate that, as the optimization error decays to 0, our proposed estimator has a statistical rate of convergence of approximately

$$\frac{\surd\lambda_1(\mathbf{A})}{\lambda_1(\mathbf{A}) - \lambda_2(\mathbf{A})}\sqrt{\left\{\frac{s\log(d)}{n}\right\}},$$

which matches the minimax optimal rate of convergence for sparse PCA problems (Cai *et al.*, 2013).

4.2.2. *Sparse canonical correlation analysis*

For sparse CCA, we assume the model

$$\begin{pmatrix}\mathbf{X}\\ \mathbf{Y}\end{pmatrix} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$$

and

$$\boldsymbol{\Sigma} = \begin{pmatrix}\boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_{xy}\\ \boldsymbol{\Sigma}_{xy}^{\mathrm{T}} & \boldsymbol{\Sigma}_y\end{pmatrix}.$$

Recall from example 2 the definitions of $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ in the context of sparse CCA. The following proposition characterizes the rate of convergence between $\hat{\boldsymbol{\Sigma}}$ and $\boldsymbol{\Sigma}$. It follows from lemma 6.5 of Gao *et al.* (2017). For ease of presentation, we omit the dependence on the eigenvalues of $\mathbf{A}$ and $\mathbf{B}$ for CCA.

Proposition 2. Let $\hat{\boldsymbol{\Sigma}}_x$, $\hat{\boldsymbol{\Sigma}}_y$ and $\hat{\boldsymbol{\Sigma}}_{xy}$ be the sample covariances of $\boldsymbol{\Sigma}_x$, $\boldsymbol{\Sigma}_y$ and $\boldsymbol{\Sigma}_{xy}$ respectively. For any $C > 0$ and positive integer $\bar{k}$, there is a constant $C' > 0$ such that

$$\rho(\hat{\boldsymbol{\Sigma}}_x - \boldsymbol{\Sigma}_x, \bar{k}) \leqslant C\sqrt{\left\{\frac{\bar{k}\log(d)}{n}\right\}},$$

$$\rho(\hat{\boldsymbol{\Sigma}}_y - \boldsymbol{\Sigma}_y, \bar{k}) \leqslant C\sqrt{\left\{\frac{\bar{k}\log(d)}{n}\right\}}$$

and

$$\rho(\hat{\boldsymbol{\Sigma}}_{xy} - \boldsymbol{\Sigma}_{xy}, \bar{k}) \leqslant C\sqrt{\left\{\frac{\bar{k}\log(d)}{n}\right\}},$$

with high probability. Moreover, $\|\hat{\boldsymbol{\Sigma}}_{xy} - \boldsymbol{\Sigma}_{xy}\|_{\infty,\infty} \leqslant C\sqrt{\{\log(d)/n\}}$ with high probability.

We now verify the sample size condition in proposition 1. From proposition 2, we have $\rho(\mathbf{E}_{\mathbf{B}}, s^2) = \mathcal{O}_P[\sqrt{\{s^2\log(d)/n\}}]$. Thus, we need $n > Cs^2\log(d)$ for some generic constant C. Under the sample size condition and using the results in proposition 2, it can be shown that $\|\tilde{\mathbf{A}} - \hat{\mathbf{A}}\|_{\infty,\infty} \leqslant \|\tilde{\mathbf{A}} - \mathbf{A}\|_{\infty,\infty} + \|\hat{\mathbf{A}} - \mathbf{A}\|_{\infty,\infty} = \mathcal{O}_P\{\sqrt{\log(d)/n}\}$. Moreover, $\|\hat{\mathbf{B}}_{\mathcal{S}_v} - \mathbf{B}_{\mathcal{S}_v}\|_2 = \mathcal{O}_P\{\sqrt{(s/n)}\}$. Thus, as long as $n > Cs^2\log(d)$, $\mathbf{v}_0$ converges to $\mathbf{v}^*$. This verifies the assumption on $\mathbf{v}_0$ in theorem 1.

Recently Ma and Li (2016) showed that the minimax optimal eigenfactor takes the form $\sqrt{(1-\lambda_1^2)}\sqrt{(1-\lambda_2^2)}/(\lambda_1 - \lambda_2)$ in the low dimensional setting in which $n > d$, under the assumption that $\boldsymbol{\Sigma}_x = \boldsymbol{\Sigma}_y = \mathbf{I}$. Adapting the results in Ma and Li (2016) in a similar fashion to that in corollary 2, theorem 1 indicates that, with high probability, our proposed estimator obtains the minimax statistical rate of convergence of approximately

$$\frac{\sqrt{(1-\lambda_1^2)}\sqrt{(1-\lambda_2^2)}}{\lambda_1 - \lambda_2}\sqrt{\left\{\frac{s\log(d)}{n}\right\}}, \tag{17}$$

for the case when $\boldsymbol{\Sigma}_x = \boldsymbol{\Sigma}_y = \mathbf{I}$. However, the minimax optimal eigenfactor for general $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_y$ remains an open problem in the literature.

To obtain the rate of convergence for general $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_y$, we shall apply corollary 1 to the sparse CCA problem. Choosing k to be of the same order as s, proposition 2 implies that both $\rho(\mathbf{E}_{\mathbf{A}}, k')$ and $\rho(\mathbf{E}_{\mathbf{B}}, k')$ are of the order of $\sqrt{\{s\log(d)/n\}}$ with high probability. Thus, corollary 1 indicates that, as the optimization error decays to 0, our proposed estimator has a statistical rate of convergence of approximately

$$\frac{\sqrt{(1+\lambda_1^2)}\sqrt{(1+\lambda_2^2)}}{\lambda_1 - \lambda_2}\sqrt{\left\{\frac{s\log(d)}{n}\right\}}. \tag{18}$$

The upper bound is expected to be loose in terms of the eigenfactor since the class of paired matrices $(\mathbf{A}, \mathbf{B})$ that was considered in corollary 1 is a much larger class of matrices than that of the sparse CCA.

In short, our theoretical results are very general and are not based on any statistical model. Moreover, the results in theorem 1 are written as a function of the estimation error between $\mathbf{v}(F)$, the solution of a GEP restricted on the set F, and $\mathbf{v}^*$. Therefore, existing minimax optimal results for various statistical models in the low dimensional setting can be adapted to the high dimensional setting in a similar fashion to that in corollary 2.

5. Numerical studies

We perform extensive numerical studies to evaluate the performance of our proposal, rifle, compared with existing methods. We consider sparse FDA and sparse CCA, each of which can be recast as the sparse GEP (2), as shown in examples 1 and 2.

Rifle involves an initial vector $\mathbf{v}_0$ and a tuning parameter k on the cardinality. We employ the convex optimization approach that was proposed in Section 3.2 to obtain an initial vector $\mathbf{v}_0$. The convex approach involves two tuning parameters: we simply select $\zeta = \surd\{\log(d)/n\}$ and $K = 1$ as suggested by the theoretical analysis. Note that these tuning parameters can be selected conservatively since there is a refinement step to obtain a final estimator by using rifle.

It is challenging to propose a general model selection technique for the selection of k in a sparse GEP since it is not based on any statistical model and it includes both unsupervised learning and supervised learning methods as its special cases. For supervised learning methods such as sparse FDA, we perform cross-validation to select the truncation parameter k. For unsupervised learning methods such as sparse PCA and CCA, it is generally agreed in the literature that the model selection problem is challenging. In principle, we could also use cross-validation techniques to select k in these settings such as the procedure that was considered in Witten *et al.* (2009). For simplicity, in our simulation studies, we assess the performance of our estimator in the context of sparse CCA across several values of k and examine the role of k under finite sample settings.

5.1. *Fisher's discriminant analysis*

We consider high dimensional classification problems using sparse FDA. The data consist of an $n \times d$ matrix $\mathbf{X}$ with d features measured on n observations, each of which belongs to one of K classes. We let $\mathbf{x}_i$ denote the ith row of $\mathbf{X}$, and let $C_k \subset \{1, \ldots, n\}$ contain the indices of the observations in the kth class with $n_k = |C_k|$ and $\Sigma_{k=1}^{K} n_k = n$.

Recall from example 1 that this is a special case of the sparse GEP with $\hat{\mathbf{A}} = \hat{\mathbf{\Sigma}}_{\mathrm{b}}$ and $\hat{\mathbf{B}} = \hat{\mathbf{\Sigma}}_{\mathrm{w}}$. Let $\hat{\boldsymbol{\mu}}_k = \Sigma_{i \in C_k} \mathbf{x}_i / n_k$ be the estimated mean for the kth class. The standard estimates for $\mathbf{\Sigma}_{\mathrm{w}}$ and $\mathbf{\Sigma}_{\mathrm{b}}$ are

$$\hat{\mathbf{\Sigma}}_{\mathrm{w}} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in C_k} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^{\mathrm{T}}$$

and

$$\hat{\mathbf{\Sigma}}_{\mathrm{b}} = \frac{1}{n} \sum_{k=1}^{K} n_k \hat{\boldsymbol{\mu}}_k \hat{\boldsymbol{\mu}}_k^{\mathrm{T}}.$$

We consider two simulation settings similar to that of Witten *et al.* (2009).

(a) Binary classification: in this example, we set $\boldsymbol{\mu}_1 = \mathbf{0}$, $\mu_{2j} = 0.5$ for $j = \{2, 4, \ldots, 40\}$ and $\mu_{2j} = 0$ otherwise. Let $\mathbf{\Sigma}$ be a block diagonal covariance matrix with five blocks, each of dimension $d/5 \times d/5$. The (j, j')th element of each block takes value $0.8^{|j-j'|}$. As suggested by Witten *et al.* (2009), this covariance structure is intended to mimic the covariance structure of gene expression data. The data are simulated as $\mathbf{x}_i \sim N(\boldsymbol{\mu}_k, \mathbf{\Sigma})$ for $i \in C_k$.

(b) Multiclass classification: there are $K = 4$ classes in this example. Let $\mu_{kj} = (k-1)/3$ for $j = \{2, 4, \ldots, 40\}$ and $\mu_{kj} = 0$ otherwise. The data are simulated as $\mathbf{x}_i \sim N(\boldsymbol{\mu}_k, \mathbf{\Sigma})$ for $i \in C_k$, with the same covariance structure for binary classification. As noted in Witten *et al.* (2009), a one-dimensional vector projection of the data fully captures the class structure.

Table 3. Number of misclassified observations out of 1000 test samples and number of non-zero features (and standard errors) for binary and multiclass classification problems, averaged over 200 data sets†

Problem		*Results for the following methods:*				
		l_1 penalized	*l_1-FDA*	*Direct*	*Rifle*	*Oracle*
Binary	Error	32 (1)	298 (1)	29 (1)	15 (1)	8 (1)
	Features	88 (1)	23 (1)	105 (2)	42 (1)	41 (0)
Multiclass	Error	495 (2)	497 (1)	247 (2)	192 (2)	153 (1)
	Features	54 (2)	22 (1)	102 (2)	42 (1)	41 (0)

†The results (rounded to the nearest integer) are for models trained with 400 training samples with 500 features.

Four approaches are compared:

(a) rifle;
(b) l_1-penalized logistic or multinomial regression implemented by using the R package `glmnet`;
(c) l_1-penalized FDA with a diagonal estimate of $\mathbf{\Sigma}_\mathrm{w}$ implemented by using the R package `penalizedLDA` (Witten *et al.*, 2009);
(d) a direct approach to sparse discriminant analysis (Mai *et al.*, 2012, 2016) implemented by using the R package `dsda` and `msda` for binary and multiclass classification respectively.

For each method, models are fitted on the training set with tuning parameter selected by using fivefold cross-validation. Then, the models are evaluated on the test set. In addition to the aforementioned models, we consider an oracle estimator using the theoretical direction $\mathbf{v}^*$, computed by using the population quantities $\mathbf{\Sigma}_\mathrm{w}$ and $\mathbf{\Sigma}_\mathrm{b}$.

To compare the performance of the various proposals, we report the misclassification error on the test set and the number of non-zero features that are selected in the models. The results for 400 training samples and 1000 test samples, with $d=500$ features, are reported in Table 3. From Table 3, we see that rifle has the lowest misclassification error compared with other competing methods. This suggests that algorithm 1 works well with the initial value that is obtained from the convex approach in Section 3.2. The method of Witten *et al.* (2009) has the highest misclassification error in both of our simulation settings, since it does not take into account the dependences between the features. The methods of Mai *et al.* (2012, 2015) perform slightly worse than our proposal in terms of misclassification error. Moreover, they used a large number of features in their model, which renders interpretation difficult. In contrast, the number of features that are selected by our proposal is very close to that of the oracle estimator.

5.2. Canonical correlation analysis

In this section, we study the relationship between two sets of random variables $\mathbf{X} \in \mathbb{R}^{d/2}$ and $\mathbf{Y} \in \mathbb{R}^{d/2}$ in the high dimensional setting using sparse CCA. Let $\mathbf{\Sigma}_x$ and $\mathbf{\Sigma}_y$ be the covariance matrices of $\mathbf{X}$ and $\mathbf{Y}$, and $\mathbf{\Sigma}_{xy}$ be the cross-covariance matrix of $\mathbf{X}$ and $\mathbf{Y}$. We consider two different scenarios in which $\mathbf{\Sigma}_{xy}$ is low rank and approximately low rank.

Throughout the simulation studies, we compare our proposal with that of Witten *et al.* (2009), implemented by using the R package PMA. Their proposal involves choosing two tuning parameters that control the sparsity of the estimated directional vectors. We consider a range of tuning parameters and choose tuning parameters that yield the lowest estimation error for Witten *et al.* (2009). We assess the performance of rifle by considering multiple values of $k=\{6,8,10,15\}$.

The output of both our proposal and that of Witten *et al.* (2009) is normalized to have norm 1, whereas the true parameters $\mathbf{v}_x^*$ and $\mathbf{v}_y^*$ are normalized with respect to $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_y$. To evaluate the performance of the two methods, we normalize $\mathbf{v}_x^*$ and $\mathbf{v}_y^*$ such that they have norm 1 and compute the squared l_2-distance between the estimated and the true directional vectors.

5.2.1. Low rank $\boldsymbol{\Sigma}_{xy}$

Assume that $(\mathbf{X},\mathbf{Y})\sim N(\mathbf{0},\boldsymbol{\Sigma})$ with

$$\boldsymbol{\Sigma}=\begin{pmatrix}\boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_{xy}\\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_y\end{pmatrix}$$

and

$$\boldsymbol{\Sigma}_{xy}=\boldsymbol{\Sigma}_x\mathbf{v}_x^*\lambda_1(\mathbf{v}_y^*)^{\mathrm{T}}\boldsymbol{\Sigma}_y,$$

where $0<\lambda_1<1$ is the largest generalized eigenvalue and $\mathbf{v}_x^*$ and $\mathbf{v}_y^*$ are the leading pair of canonical directions. The data consists of two $n\times d/2$ matrices $\mathbf{X}$ and $\mathbf{Y}$. We assume that each row of the two matrices is generated according to $(\mathbf{x}_i,\mathbf{y}_i)\sim N(\mathbf{0},\boldsymbol{\Sigma})$. The goal of CCA is to estimate the canonical directions $\mathbf{v}_x^*$ and $\mathbf{v}_y^*$ on the basis of the data matrices $\mathbf{X}$ and $\mathbf{Y}$.

Let $\hat{\boldsymbol{\Sigma}}_x$ and $\hat{\boldsymbol{\Sigma}}_y$ be the sample covariance matrices of $\mathbf{X}$ and $\mathbf{Y}$, and let $\hat{\boldsymbol{\Sigma}}_{xy}$ be the sample cross-covariance matrix of $\mathbf{X}$ and $\mathbf{Y}$. Recall from example 2 that the sparse CCA problem can be recast as the GEP with

$$\hat{\mathbf{A}}=\begin{pmatrix}\mathbf{0} & \hat{\boldsymbol{\Sigma}}_{xy}\\ \hat{\boldsymbol{\Sigma}}_{xy} & \mathbf{0}\end{pmatrix},$$

$$\hat{\mathbf{B}}=\begin{pmatrix}\hat{\boldsymbol{\Sigma}}_x & \mathbf{0}\\ \mathbf{0} & \hat{\boldsymbol{\Sigma}}_y\end{pmatrix}$$

and

$$\mathbf{v}=\begin{pmatrix}\mathbf{v}_x\\ \mathbf{v}_y\end{pmatrix}.$$

In our simulation setting, we set $\lambda_1=0.9$, $v_{x,j}^*=v_{y,j}^*=1/\sqrt{3}$ for $j=\{1,6,11\}$, and $v_{x,j}^*=v_{y,j}^*=0$ otherwise. Then, we normalize $\mathbf{v}_x^*$ and $\mathbf{v}_y^*$ such that $(\mathbf{v}_x^*)^{\mathrm{T}}\boldsymbol{\Sigma}_x\mathbf{v}_x^*=(\mathbf{v}_y^*)^{\mathrm{T}}\boldsymbol{\Sigma}_y\mathbf{v}_y^*=1$. We consider the case when $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_y$ are block diagonal matrices with five blocks, each of dimension $d/5\times d/5$, where the (j,j')th element of each block takes value $0.8^{|j-j'|}$. The results for $d=500$ and $s=6$, averaged over 200 data sets, are summarized in Table 4.

From Table 4, we see that our proposal outperforms that of Witten *et al.* (2009) uniformly across different sample sizes. This is not surprising since Witten *et al.* (2009) used diagonal estimates of $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_y$ to compute the directional vectors. The l_2-distance for our proposal decreases as we increase n. Moreover, the l_2-distance increases when we increase k. These results confirm our theoretical analysis in theorem 1.

Table 4. Results for low rank Σ_{xy}†

n	*Results for the following methods:*				
	PMA	*Rifle (k=6)*	*Rifle (k=8)*	*Rifle (k=10)*	*Rifle (k=15)*
$\mathbf{v}_x$					
200	0.72 (0.01)	0.21 (0.02)	0.11 (0.02)	0.08 (0.02)	0.07 (0.01)
400	0.61 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
600	0.58 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
$\mathbf{v}_y$					
200	0.70 (0.01)	0.24 (0.02)	0.24 (0.02)	0.35 (0.02)	0.58 (0.01)
400	0.62 (0.01)	0.02 (0.01)	0.07 (0.01)	0.15 (0.01)	0.32 (0.01)
600	0.59 (0.01)	0.01 (0.01)	0.04 (0.01)	0.08 (0.01)	0.19 (0.01)

†Squared l_2-distance between the estimated and true leading generalized eigenvector as a function of the sample size n for $d=500$ and $s=6$. The results are averaged over 200 data sets.

5.2.2. *Approximately low rank Σ_{xy}*

In this section, we consider the case when Σ_{xy} is approximately low rank. We consider the same simulation set-up as in the previous section, except that Σ_{xy} is now approximately low rank, generated as follows:

$$\Sigma_{xy}=\Sigma_x\mathbf{v}_x^*\lambda_1(\mathbf{v}_y^*)^{\mathrm{T}}\Sigma_y+\Sigma_x\mathbf{V}_x^*\boldsymbol{\Lambda}(\mathbf{V}_y^*)^{\mathrm{T}}\Sigma_y$$

with $\lambda_1=0.9$. Here, $\boldsymbol{\Lambda}\in\mathbb{R}^{200\times200}$ is a diagonal matrix with diagonal entries 0.1, and $\mathbf{V}_x^*,\mathbf{V}_y^*\in\mathbb{R}^{d/2\times200}$ are normalized orthogonal matrices such that $(\mathbf{V}_x^*)^{\mathrm{T}}\Sigma_x\mathbf{V}_x^*=\mathbf{I}$ and $(\mathbf{V}_y^*)^{\mathrm{T}}\Sigma_y\mathbf{V}_y^*=\mathbf{I}$ respectively. The goal is to recover the leading generalized eigenvector $\mathbf{v}_x^*$ and $\mathbf{v}_y^*$. The results for $d=1000$ and $s=6$, averaged over 200 data sets, are summarized in Table 5.

From Table 5, we see that the performance of rifle is much better than that of PMA across all settings. As we increase the number of samples n, the l_2-distance decreases for all values of k. Interesting, as we increase k from $k=6$ to $k=10$ for the case when $n=400$, the l_2-distance decreases

Table 5. Results for approximately low rank Σ_{xy}†

n	*Results for the following methods:*				
	PMA	*Rifle (k=6)*	*Rifle (k=8)*	*Rifle (k=10)*	*Rifle (k=15)*
$\mathbf{v}_x$					
400	0.63 (0.01)	0.30 (0.02)	0.19 (0.02)	0.13 (0.02)	0.07 (0.01)
600	0.62 (0.01)	0.11 (0.01)	0.07 (0.01)	0.09 (0.01)	0.07 (0.01)
800	0.57 (0.01)	0.02 (0.01)	0.05 (0.01)	0.08 (0.01)	0.07 (0.01)
$\mathbf{v}_y$					
400	0.66 (0.01)	0.31 (0.02)	0.26 (0.02)	0.22 (0.02)	0.25 (0.01)
600	0.63 (0.01)	0.10 (0.01)	0.11 (0.01)	0.13 (0.01)	0.16 (0.01)
800	0.55 (0.01)	0.02 (0.01)	0.07 (0.01)	0.11 (0.01)	0.13 (0.01)

†Squared l_2-distance between the estimated and true leading generalized eigenvector as a function of the sample size n for $d=1000$ and $s=6$. The results are averaged over 200 data sets.

slightly. This is because, in the high dimensional setting, the initial value is not estimated accurately. Thus, when we choose $k=s=6$, some of the true support is not selected after truncating the initial value $\mathbf{v}_0$ and therefore it has a higher l_2-distance. In this case, by selecting a larger value of k, we can ensure that the true support is selected, which yields a lower l_2-distance. If an even larger k is selected, then the l_2-distance will eventually increase like in the case when $k=15$ for $\mathbf{v}_y$.

6. Data application

In this section, we apply our method in the context of sparse sliced inverse regression as in example 3. The data sets that we consider are as follows.

(a) The leukaemia (Golub *et al.*, 1999) data set consists of 7129 gene expression measurements from 25 patients with acute myeloid leukaemia and 47 patients with acute lymphoblastic leukaemia. The data are available from `http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi`. Recently, this data set has been analysed in the context of sparse sufficient dimension reduction in Yin and Hilafu (2015).

(b) The lung cancer (Spira *et al.*, 2007) data set consists of 22283 gene expression measurements from large airway epithelial cells sampled from 97 smokers with lung cancer and 90 smokers without lung cancer. The data are publicly available from the gene expression omnibus, accession number GDS2771.

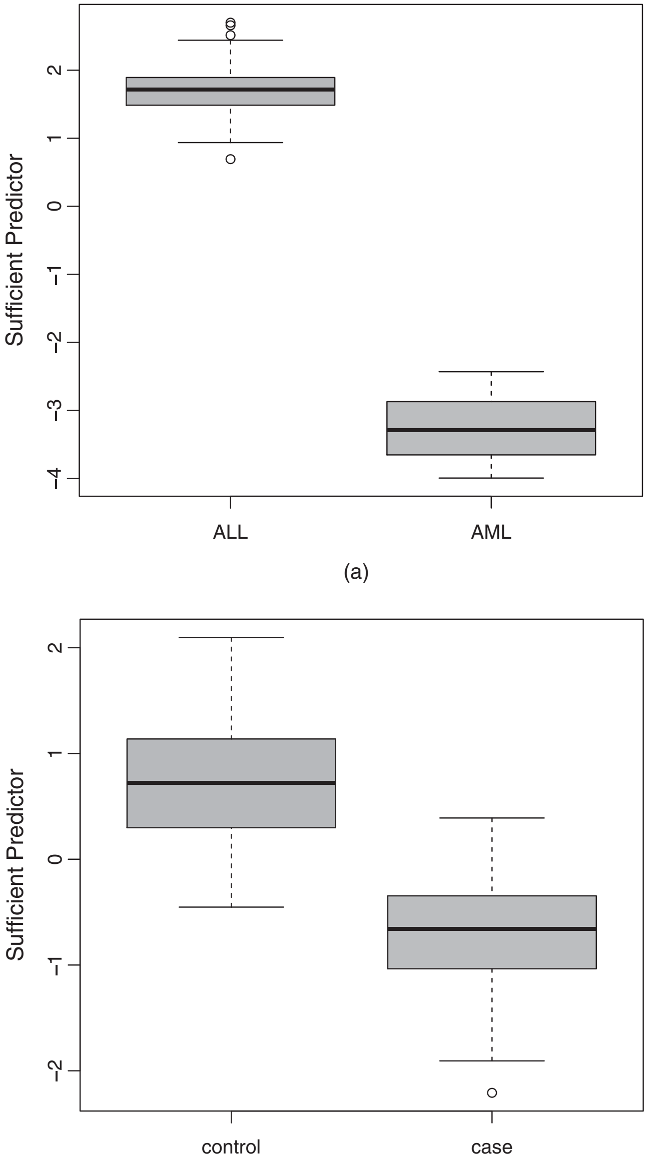
We preprocess the leukaemia data set following Golub *et al.* (1999) and Yin and Hilafu (2015). In particular, we set gene expression readings of 100 or fewer to 100, and expression readings of 16000 or more to 16000. We then remove genes with difference and ratio between the maximum and minimum readings that are less than 500 and 5 respectively. A log-transformation is then applied to the data. This gives us a data matrix $\mathbf{X}$ with 72 rows or samples and 3571 columns or genes. For the lung cancer data, we simply select the 2000 genes with the largest variance as in Petersen *et al.* (2016). This gives a data matrix with 167 rows or samples and 2000 columns or genes. We further standardize both data sets so that the genes have mean 0 and variance 1.

Recall from example 3 that, to apply our method, we need the estimates $\hat{\mathbf{A}}=\hat{\mathbf{\Sigma}}_{E(\mathbf{X}|Y)}$ and $\hat{\mathbf{B}}=\hat{\mathbf{\Sigma}}_x$. The quantity $\hat{\mathbf{\Sigma}}_x$ is simply the sample covariance matrix of $\mathbf{X}$. Let n_1 and n_2 be the number of samples of the two classes in the data set. Let $\hat{\mathbf{\Sigma}}_{x,1}$ and $\hat{\mathbf{\Sigma}}_{x,2}$ be the sample covariance matrix calculated by using only data from class 1 and class 2 respectively. Then, the covariance matrix of the conditional expectation can be estimated by

$$\hat{\mathbf{\Sigma}}_{E(\mathbf{X}|Y)}=\hat{\mathbf{\Sigma}}_x-\frac{1}{n}\sum_{k=1}^{2}n_k\hat{\mathbf{\Sigma}}_{x,k},$$

where $n=n_1+n_2$ (Li, 1991; Li and Nachtsheim, 2006; Zhu *et al.*, 2006; Li and Yin, 2008; Chen *et al.*, 2010; Yin and Hilafu, 2015). Let $\hat{\mathbf{v}}_t$ be the output of algorithm 1. Similarly to Yin and Hilafu (2015), we plot the boxplot of the sufficient predictor $\mathbf{X}\hat{\mathbf{v}}_t$ for the two classes in each data set. The results with $k=25$ for the leukaemia and lung cancer data sets are in Figs 1(a) and 1(b) respectively.

From Fig. 1(a), for the leukaemia data set, we see that the sufficient predictors for the two groups are much more well separated than the results in Yin and Hilafu (2015). Moreover, our proposal is with theoretical guarantees whereas their proposal is sequential without theoretical guarantees. For the lung cancer data set, we see that there is some overlap between the sufficient predictors for subjects with and without lung cancer. These results are consistent with the literature where it is known that the lung cancer data set is a much more difficult classification



(a)

(b)

Fig. 1. Boxplots of the sufficient predictor $\mathbf{X}\hat{\mathbf{v}}_t$ obtained from algorithm 1 for the leukaemia and lung cancer data sets: patients with acute lymphoblastic leukaemia (ALL) and acute myeloid leukaemia (AML); (b) patients with (case) and without (control) lung cancer

problem compared with that of the leukaemia data set (Fan and Fan, 2008; Petersen *et al.*, 2016).

7. Discussion

We propose a two-stage computational framework for solving the sparse GEP. The method successfully handles ill-conditioned normalization matrices that arise from the high dimensional

Table 6. Estimation error between the true standardized generalized eigenvector $\|\mathbf{v}^*\|_2 = 1$ and the estimated generalized eigenvector for the binary classification problem, averaged over 50 data sets†

	Results for soft-rifle			*Results for rifle*		
	C = 1	*C = 0.5*	*C = 0.25*	*k = 35*	*k = 40*	*k = 55*
Estimation error	0.180	0.048	0.072	0.181	0.048	0.072
Features	33.5	39.7	53.3	35	40	55

†The numbers of non-zero features are also reported. The results are with $n = 200$ and $d = 200$. The true sparsity level is $s = 40$.

setting because of finite sample estimation, and the final estimator enjoys geometric convergence to a solution with the optimal statistical rate of convergence. Our method and theory have applications to a large class of statistical models including but not limited to sparse FDA, sparse CCA and sparse sufficient dimension reduction. Compared with existing theory for each specific statistical model, our theory is very general and does not require any structural assumptions on $(\mathbf{A}, \mathbf{B})$.

Our theoretical results in theorem 1 rely on selecting the tuning parameter k such that $k = Cs$ for some constant $C > 1$. However, in practice, the true sparsity level s is unknown and it may be difficult to select the value of k. To remove the dependences on s, one of the reviewers suggested a thresholding strategy, i.e. instead of truncating the vector $\mathbf{v}_t'$ and keeping the top k elements, one can perform $C\sqrt{\{\log(d)/n\}}$ thresholding on the updated vector $\mathbf{v}_t'$ from step 3 of algorithm 1, where C is some user-specified constant. To evaluate the thresholding strategy, we perform a small scale numerical study on the FDA binary classification example similar to that of Section 5.1 with $n = 200$ and $d = 200$. We compare the estimator that is obtained by using the soft thresholding rule, soft-rifle, and that of our proposed truncation rule by calculating the estimation error between these estimators and the oracle direction. The results, averaged across 50 iterations, are presented in Table 6. From Table 6, we see that, depending on the choice of the constant C, the soft thresholding rule has a similar performance to that of the truncation rule, suggesting that substituting the soft thresholding rule in steps 4 and 5 of algorithm 1 will also work.

In the case when $\mathbf{v}^*$ is approximately sparse, i.e. $s = d$, the current theoretical results are no longer applicable. To address this issue, we can redefine the notion of sparsity level s. As suggested by one of the reviewers, we can define the effective sparsity level s' as the l_q-norm $(q < 1)$ or the ratio between, for example, l_1- and l_∞-norms of $\mathbf{v}^*$. The theoretical properties for the thresholding strategy and weak sparsity are challenging to establish under our current theoretical framework. In particular, because of the normalization constraint $\mathbf{v}^{\mathrm{T}}\hat{\mathbf{B}}\mathbf{v}$ on the denominator, to analyse the gradient ascent step in step 2, we require that the cardinality of the input vector must have support k'. This condition is needed to control the condition number of $\hat{\mathbf{B}}_F$, where F is an index set such that $|F| = k'$. Developing a new theoretical framework for solving the sparse generalized eigenvalue problem is out of the scope of this paper and we leave it for future work.

There are several additional future directions for the sparse GEP. It will be interesting to study whether rifle can be generalized to the case for estimating subspace spanned by the top K leading generalized eigenvectors. The computational bottleneck for the current approach is on the convex relaxation method for obtaining the initial vector $\mathbf{v}_0$, which has a computational complexity of $\mathcal{O}(d^3)$ per iteration. This yields a total computational complexity of $\mathcal{O}(d^3) + \mathcal{O}(kd + d)$ for the proposed two-stage computational framework. In future work, it will be of paramount importance to propose an efficient convex algorithm to obtain $\mathbf{v}_0$ such that our proposal is scalable to accommodate large-scale data.

Acknowledgements

We thank the Joint Editor, Associate Editor and two reviewers for their helpful comments that improved an earlier version of this paper. We thank Gao Chao and Xiaodong Li for responding to our enquiries. Tong Zhang was supported by National Science Foundation grants NSF IIS-1250985 and NSF IIS-1407939 and National Institutes of Health grant NIH R01AI116744. Kean Ming Tan was supported by National Science Foundation grants NSF IIS-1250985, NSF IIS-1407939 and NSF DMS-1811315.

Appendix A: Proof of theorem 1

To establish theorem 1, we first quantify the error that is introduced by maximizing the empirical version of the GEP, restricted to a superset of V ($V \subset F$), i.e.

$$\mathbf{v}(F) = \underset{\mathbf{v}\in\mathbb{R}^d}{\arg\max}\ \mathbf{v}^{\mathrm{T}}\hat{\mathbf{A}}\mathbf{v}, \qquad \text{subject to } \mathbf{v}^{\mathrm{T}}\hat{\mathbf{B}}\mathbf{v}=1, \quad \mathrm{supp}(\mathbf{v}) \subseteq F.$$

Then we establish an error bound between $\mathbf{v}_t'$ in step 2 of algorithm 1 and $\mathbf{v}(F)$. Finally, we quantify the error that is introduced by the truncated step in algorithm 1.

We first state a series of lemmas that will facilitate the proof of theorem 1. The proofs for the technical lemmas are deferred to Appendix C. We start with some results from perturbation theory for eigenvalues and GEPs (Golub and Van Loan, 2012).

Lemma 1. Let $\mathbf{J}$ and $\mathbf{J}+\mathbf{E}_\mathbf{J}$ be $d \times d$ symmetric matrices. Then, for all $k \in \{1,\ldots,d\}$,

$$\lambda_k(\mathbf{J}) + \lambda_{\min}(\mathbf{E}_\mathbf{J}) \leqslant \lambda_k(\mathbf{J}+\mathbf{E}_\mathbf{J}) \leqslant \lambda_k(\mathbf{J}) + \lambda_{\max}(\mathbf{E}_\mathbf{J}).$$

In what follows, we state a result on perturbed generalized eigenvalues for a symmetric definite matrix pair $(\mathbf{J},\mathbf{K})$ in the following lemma, which follows directly from theorem 3.2 in Stewart (1979) and theorem 8.7.3 in Golub and Van Loan (2012).

Lemma 2. Let $(\mathbf{J},\mathbf{K})$ be a symmetric definite matrix pair with generalized eigenvalues $\lambda_1(\mathbf{J},\mathbf{K}) \geqslant \ldots \geqslant \lambda_d(\mathbf{J},\mathbf{K})$. Let $(\mathbf{J}+\mathbf{E}_\mathbf{J},\mathbf{K}+\mathbf{E}_\mathbf{K})$ be the perturbed matrix pair and assume that $\mathbf{E}_\mathbf{J}$ and $\mathbf{E}_\mathbf{K}$ satisfy

$$\epsilon = \surd(\|\mathbf{E}_\mathbf{J}\|_2^2 + \|\mathbf{E}_\mathbf{K}\|_2^2) < \mathrm{cr}(\mathbf{J},\mathbf{K}),$$

where $\mathrm{cr}(\mathbf{J},\mathbf{K})$ is as defined in expression (10). Then, $(\mathbf{J}+\mathbf{E}_\mathbf{J},\ \mathbf{K}+\mathbf{E}_\mathbf{K})$ is a symmetric definite matrix pair with generalized eigenvalues $\lambda_1(\mathbf{J}+\mathbf{E}_\mathbf{J},\mathbf{K}+\mathbf{E}_\mathbf{K}) \geqslant \ldots \geqslant \lambda_d(\mathbf{J}+\mathbf{E}_\mathbf{J},\mathbf{K}+\mathbf{E}_\mathbf{K})$. Then,

$$\frac{\lambda_k(\mathbf{J},\mathbf{K})\,\mathrm{cr}(\mathbf{J},\mathbf{K}) - \epsilon}{\mathrm{cr}(\mathbf{J},\mathbf{K}) + \epsilon\,\lambda_k(\mathbf{J},\mathbf{K})} \leqslant \lambda_k(\mathbf{J}+\mathbf{E}_\mathbf{J},\mathbf{K}+\mathbf{E}_\mathbf{K}) \leqslant \frac{\lambda_k(\mathbf{J},\mathbf{K})\,\mathrm{cr}(\mathbf{J},\mathbf{K}) + \epsilon}{\mathrm{cr}(\mathbf{J},\mathbf{K}) - \epsilon\,\lambda_k(\mathbf{J},\mathbf{K})}.$$

Recall from Section 4 that $\mathbf{v}^*$ is the first generalized eigenvector of $(\mathbf{A},\mathbf{B})$ with generalized eigenvalue λ_1, and that $V = \mathrm{supp}(\mathbf{v}^*)$. For any given set F such that $V \subset F$, let $\lambda_k(F)$ and $\hat{\lambda}_k(F)$ be the kth generalized eigenvalues of $(\mathbf{A}_F,\mathbf{B}_F)$ and $(\hat{\mathbf{A}}_F,\hat{\mathbf{B}}_F)$ respectively. Under assumption 1 and by application of lemma 2, we have

$$\hat{\lambda}_2(F)\ /\hat{\lambda}_1(F) \leqslant \gamma,$$

where $\gamma = (1+a)\lambda_2/\{(1-a)\lambda_1\}$.

Let $\mathbf{y}(F) = \mathbf{v}(F)/\|\mathbf{v}(F)\|_2$ and $\mathbf{y}^* = \mathbf{v}^*/\|\mathbf{v}^*\|_2$ such that $\|\mathbf{y}(F)\|_2 = \|\mathbf{y}^*\|_2 = 1$. We now present a key lemma on measuring the progress of the gradient descent step. It requires an initial solution that is sufficiently close to the optimal value in expression (14). With some abuse of notation, we indicate $\mathbf{y}(F)$ to be a k'-dimensional vector restricted to the set $F \subset \{1,\ldots,d\}$ with $|F| = k'$. Recall that $c > 0$ is some arbitrary small constant stated in assumption 1 and c_{upper} is defined as $(1+c)/(1-c)$.

Lemma 3. Let $F \subset \{1,\ldots,d\}$ be some set with $|F| = k'$. Given any $\tilde{\mathbf{v}}$ such that $\|\tilde{\mathbf{v}}\|_2 = 1$ and $\tilde{\mathbf{v}}^{\mathrm{T}}\mathbf{y}(F) > 0$, let $\rho = \tilde{\mathbf{v}}^{\mathrm{T}}\hat{\mathbf{A}}_F\tilde{\mathbf{v}}/\tilde{\mathbf{v}}^{\mathrm{T}}\hat{\mathbf{B}}_F\tilde{\mathbf{v}}$, and let $\mathbf{v}' = \mathbf{C}_F\tilde{\mathbf{v}}/\|\mathbf{C}_F\tilde{\mathbf{v}}\|_2$, where

$$\mathbf{C} = \mathbf{I} + (\eta/\rho)(\hat{\mathbf{A}} - \rho\hat{\mathbf{B}})$$

and $\eta > 0$ is some positive constant. Let $\delta = 1 - \mathbf{y}(F)^{\mathrm{T}}\tilde{\mathbf{v}}$. Pick η sufficiently small such that

$$\eta\,\lambda_{\max}(\mathbf{B}) < 1/(1+c),$$

and δ is sufficiently small such that

$$1 - \delta \geqslant 1 - \theta(\mathbf{A},\mathbf{B}),$$

where

$$\theta(\mathbf{A},\mathbf{B}) = \min\left[\frac{1}{8c_{\text{upper}}\kappa(\mathbf{B})}, \frac{1/\gamma - 1}{3c_{\text{upper}}\kappa(\mathbf{B})}, \frac{1-\gamma}{30(1+c)c_{\text{upper}}^2\eta\,\lambda_{\max}(\mathbf{B})\,\kappa^2(\mathbf{B})\{c_{\text{upper}}\kappa(\mathbf{B}) + \gamma\}}\right].$$

Then, under assumption 1, we have

$$\mathbf{y}(F)^{\mathrm{T}}\mathbf{v}' \geqslant \mathbf{y}(F)^{\mathrm{T}}\tilde{\mathbf{v}} + \frac{1+c}{8}\eta\,\lambda_{\min}(\mathbf{B})\{1 - \mathbf{y}(F)^{\mathrm{T}}\tilde{\mathbf{v}}\}\left\{\frac{1-\gamma}{c_{\text{upper}}\kappa(\mathbf{B}) + \gamma}\right\}.$$

The following lemma characterizes the error that is introduced by the truncation step. It follows directly from lemma 12 in Yuan and Zhang (2013).

Lemma 4. Consider $\mathbf{y}'$ with $F' = \operatorname{supp}(\mathbf{y}')$ and $|F'| = \bar{k}$. Let F be the indices of $\mathbf{y}$ with the largest k absolute values, with $|F| = k$. If $\|\mathbf{y}'\|_2 = \|\mathbf{y}\|_2 = 1$, then

$$|\text{truncate}(\mathbf{y}, F)^{\mathrm{T}}\mathbf{y}'| \geqslant |\mathbf{y}^{\mathrm{T}}\mathbf{y}'| - (\bar{k}/k)^{1/2}\min[\surd\{1 - (\mathbf{y}^{\mathrm{T}}\mathbf{y}')^2\}, \{1 + (\bar{k}/k)^{1/2}\}\{1 - (\mathbf{y}^{\mathrm{T}}\mathbf{y}')^2\}].$$

Recall from algorithm 1 that we define $\mathbf{v}_t = \hat{\mathbf{v}}_t/\|\hat{\mathbf{v}}_t\|_2$. Since $\|\mathbf{v}_t'\|_2 = 1$, and $\hat{\mathbf{v}}_t$ is the truncated version of $\mathbf{v}_t'$, we have that $\|\hat{\mathbf{v}}_t\|_2 \leqslant 1$. This implies that $|(\mathbf{y}^*)^{\mathrm{T}}\mathbf{v}_t| \geqslant |(\mathbf{y}^*)^{\mathrm{T}}\hat{\mathbf{v}}_t|$. We now quantify the progress of each iteration of algorithm 1. For this, assume that $k > s$, where s is the cardinality of the support of $\mathbf{y}^* = \mathbf{v}^*/\|\mathbf{v}^*\|_2^2$, and k is the truncation parameter in algorithm 1. Let $k' = 2k + s$ and let

$$\nu = \surd[1 + 2\{(s/k)^{1/2} + s/k\}]\sqrt{\left\{1 - \frac{1+c}{8}\eta\,\lambda_{\min}(\mathbf{B})\frac{1-\gamma}{c_{\text{upper}}\kappa(\mathbf{B}) + \gamma}\right\}}.$$

Recall that V is the support of $\mathbf{v}^*$, the population leading generalized vector, and also $\mathbf{y}^* = \mathbf{v}^*/\|\mathbf{v}^*\|_2$. Let $F_{t-1} = \operatorname{supp}(\mathbf{v}_{t-1})$ and $F_t = \operatorname{supp}(\mathbf{v}_t)$ and let $F = F_{t-1} \cup F_t \cup V$. Note that the cardinality of F is no more than $k' = 2k + s$, since $|F_t| = |F_{t-1}| = k$. Let

$$\mathbf{v}_t' = \mathbf{C}_F\mathbf{v}_{t-1}/\|\mathbf{C}_F\mathbf{v}_{t-1}\|_2,$$

where $\mathbf{C}_F$ is the submatrix of $\mathbf{C}_F$ restricted to the rows and columns that are indexed by F. We note that $\mathbf{v}_t'$ is equivalent to that in algorithm 1, since the elements of $\mathbf{v}_t'$ outside the set F take value 0. Without loss of generality and for simplicity, we assume that the inner product between two eigenvectors is positive, because otherwise we can simply do appropriate sign changes in the proof.

Applying lemma 3 with the set F, we obtain

$$\mathbf{y}(F)^{\mathrm{T}}\mathbf{v}_t' \geqslant \mathbf{y}(F)^{\mathrm{T}}\mathbf{v}_{t-1} + \frac{1+c}{8}\eta\,\lambda_{\min}(\mathbf{B})\{1 - \mathbf{y}(F)^{\mathrm{T}}\mathbf{v}_{t-1}\}\,\frac{1-\gamma}{c_{\text{upper}}\kappa(\mathbf{B}) + \gamma}.$$

Subtracting 1 from both sides of the equation and rearranging the terms, we obtain

$$1 - \mathbf{y}(F)^{\mathrm{T}}\mathbf{v}_t' \leqslant \{1 - \mathbf{y}(F)^{\mathrm{T}}\mathbf{v}_{t-1}\}\left\{1 - \frac{1+c}{8}\eta\,\lambda_{\min}(\mathbf{B})\frac{1-\gamma}{c_{\text{upper}}\kappa(\mathbf{B}) + \gamma}\right\}. \tag{19}$$

This implies that

$$\|\mathbf{y}(F) - \mathbf{v}_t'\|_2 \leqslant \|\mathbf{y}(F) - \mathbf{v}_{t-1}\|_2\sqrt{\left\{1 - \frac{1+c}{8}\eta\,\lambda_{\min}(\mathbf{B})\frac{1-\gamma}{c_{\text{upper}}\kappa(\mathbf{B}) + \gamma}\right\}}. \tag{20}$$

By the triangle inequality, we have

$$\begin{aligned}\|\mathbf{y}-\mathbf{v}'_t\|_2 &\leqslant \|\mathbf{y}(F)-\mathbf{v}'_t\|_2+\|\mathbf{y}(F)-\mathbf{y}^*\|_2\\ &\leqslant \|\mathbf{y}(F)-\mathbf{v}_{t-1}\|_2\sqrt{\left\{1-\frac{1+c}{8}\eta\,\lambda_{\min}(\mathbf{B})\frac{1-\gamma}{c_{\text{upper}}\kappa(\mathbf{B})+\gamma}\right\}}+\|\mathbf{y}(F)-\mathbf{y}^*\|_2\\ &\leqslant \|\mathbf{y}-\mathbf{v}_{t-1}\|_2\sqrt{\left\{1-\frac{1+c}{8}\eta\,\lambda_{\min}(\mathbf{B})\frac{1-\gamma}{c_{\text{upper}}\kappa(\mathbf{B})+\gamma}\right\}}+2\|\mathbf{y}(F)-\mathbf{y}^*\|_2, \end{aligned} \qquad (21)$$

where the second inequality follows from inequality (19). This is equivalent to

$$\surd(1-|\mathbf{y}^{\mathrm{T}}\mathbf{v}'_t|)\leqslant\surd(1-|\mathbf{y}^{\mathrm{T}}\mathbf{v}_{t-1}|)\sqrt{\left\{1-\frac{1+c}{8}\eta\,\lambda_{\min}(\mathbf{B})\frac{1-\gamma}{c_{\text{upper}}\kappa(\mathbf{B})+\gamma}\right\}}+2\surd\{1-|\mathbf{y}(F)^{\mathrm{T}}\mathbf{y}^*|\}. \qquad (22)$$

We define

$$\nu=\surd[1+2\{(s/k)^{1/2}+s/k\}]\sqrt{\left\{1-\frac{1+c}{8}\eta\,\lambda_{\min}(\mathbf{B})\frac{1-\gamma}{c_{\text{upper}}\kappa(\mathbf{B})+\gamma}\right\}}.$$

By lemma 4 and picking $k>s$, we have

$$\begin{aligned}\surd(1-|\mathbf{y}^{\mathrm{T}}\hat{\mathbf{v}}_t|)&\leqslant\surd[1-|\mathbf{y}^{\mathrm{T}}\mathbf{v}'_t|+\{(s/k)^{1/2}+s/k\}(1-|\mathbf{y}^{\mathrm{T}}\mathbf{v}'_t|^2)]\\ &\leqslant\surd(1-|\mathbf{y}^{\mathrm{T}}\mathbf{v}'_t|)\surd[1+\{(s/k)^{1/2}+s/k\}(1+|\mathbf{y}^{\mathrm{T}}\mathbf{v}'_t|)]\\ &\leqslant\surd(1-|\mathbf{y}^{\mathrm{T}}\mathbf{v}'_t|)\surd[1+2\{(s/k)^{1/2}+s/k\}]\\ &\leqslant\nu\surd(1-|\mathbf{y}^{\mathrm{T}}\mathbf{v}_{t-1}|)+\surd20\surd\{1-|\mathbf{y}(F)^{\mathrm{T}}\mathbf{y}^*|\},\end{aligned} \qquad (23)$$

where the third inequality holds by using the fact that $|\mathbf{y}^{\mathrm{T}}\mathbf{v}'_t|\leqslant 1$, and the last inequality holds by inequality (22).

Finally, we have

$$\begin{aligned}\surd\{1-|(\mathbf{y}^*)^{\mathrm{T}}\mathbf{v}_t|\}&\leqslant\surd\{1-|(\mathbf{y}^*)^{\mathrm{T}}\hat{\mathbf{v}}_t|\}\\ &\leqslant\nu\surd\{1-|(\mathbf{y}^*)^{\mathrm{T}}\mathbf{v}_{t-1}|\}+\surd20\surd\{1-|\mathbf{y}(F)^{\mathrm{T}}\mathbf{y}^*|\}.\end{aligned} \qquad (24)$$

By recursively applying inequality (23), we have, for all $t\geqslant 0$,

$$\surd\{1-|(\mathbf{y}^*)^{\mathrm{T}}\mathbf{v}_t|\}\leqslant\nu^t\surd\{1-|(\mathbf{y}^*)^{\mathrm{T}}\mathbf{v}_0|\}+\surd20\surd\{1-|\mathbf{y}(F)^{\mathrm{T}}\mathbf{y}^*|\}/(1-\nu),$$

as desired.

Appendix B: Proof of corollary 1

Let $F\supset V$ be a superset of the support of $\mathbf{y}^*$. Recall that $\mathbf{y}(F)=\mathbf{v}(F)/\|\mathbf{v}(F)\|_2$ and $\mathbf{y}^*=\mathbf{y}^*/\|\mathbf{y}^*\|_2$. We first prove that $\mathbf{y}(F)$ is close to $\mathbf{y}^*$ for a general class of symmetric definite matrix pairs $(\mathbf{A},\mathbf{B})$. For this, we present the following lemma resulting from theorem 4.3 in Stewart (1979).

Lemma 5. Let F be a set such that $V\subset F$ with $|F|=k'>s$ and let

$$\delta(F)=\surd(\|\mathbf{E}_{\mathbf{A},F}\|_2^2+\|\mathbf{E}_{\mathbf{B},F}\|_2^2).$$

Let

$$\chi\{\lambda_1(F),\hat{\lambda}_k(F)\}=\frac{|\lambda_1(F)-\hat{\lambda}_k(F)|}{\surd\{1+\lambda_1(F)^2\}\surd\{1+\hat{\lambda}_k(F)^2\}},$$
$$\Delta\hat{\lambda}(F)=\min_{k>1}\chi\{\lambda_1(F),\hat{\lambda}_k(F)\}>0.$$

If $\delta(F)/\Delta\hat{\lambda}(F)<\mathrm{cr}(\hat{\mathbf{A}}_F,\hat{\mathbf{B}}_F)$, then

$$\frac{\min\{\|\mathbf{v}(F)-\mathbf{v}^*\|_2,\mathbf{v}(F)+\mathbf{v}^*\|_2\}}{\|\mathbf{v}^*\|_2}\leqslant\frac{\delta(F)}{\Delta\hat{\lambda}(F)\,\mathrm{cr}(\hat{\mathbf{A}}_F,\hat{\mathbf{B}}_F)}.$$

This implies that

$$\min\{\|\mathbf{y}(F)-\mathbf{y}^*\|_2, \|\mathbf{y}(F)+\mathbf{y}^*\|_2\} \leqslant \frac{2}{\Delta\lambda\{\mathrm{cr}(k')-\epsilon(k')\}}\epsilon(k'),$$

where $\Delta\lambda$, $\mathrm{cr}(k')$ and $\epsilon(k')$ are as defined in expressions (16) and (12).

By lemma 5, we have

$$\sqrt{\left\{1-\frac{|(\mathbf{v}^*)^{\mathrm{T}}\mathbf{v}_t|}{\|\mathbf{v}^*\|_2}\right\}} \leqslant \frac{2^{1/2}}{\Delta\lambda\{\mathrm{cr}(k')-\epsilon(k')\}}\epsilon(k').$$

Substituting the above inequality into theorem 1 yields the results in corollary 1.

Appendix C: Proof of technical lemmas

C.1. Proof of lemma 3

Recall that $F\subset\{1,\ldots,d\}$ is some set with cardinality $|F|=k'$. Also, recall that $\mathbf{y}(F)$ is proportional to the largest generalized eigenvector of $(\hat{\mathbf{A}}_F,\hat{\mathbf{B}}_F)$. Throughout the proof, we write $\hat{\kappa}$ to denote $\kappa(\hat{\mathbf{B}}_F)$ for notational convenience. In addition, we use the notation $\|\mathbf{v}\|^2_{\hat{\mathbf{B}}_F}$ to indicate $\mathbf{v}^{\mathrm{T}}\hat{\mathbf{B}}_F\mathbf{v}$.

Let $\boldsymbol{\xi}_j$ be the jth generalized eigenvector of $(\hat{\mathbf{A}}_F,\hat{\mathbf{B}}_F)$ corresponding to $\hat{\lambda}_j(F)$ such that

$$\boldsymbol{\xi}_j^{\mathrm{T}}\hat{\mathbf{B}}_F\boldsymbol{\xi}_k=\begin{cases}1 & \text{if } j=k,\\ 0 & \text{if } j\neq k.\end{cases}$$

Assume that $\tilde{\mathbf{v}}=\Sigma_{j=1}^{k'}\alpha_j\boldsymbol{\xi}_j$ and by definition we have $\mathbf{y}(F)=\boldsymbol{\xi}_1/\|\boldsymbol{\xi}_1\|_2$. By assumption, we have $\mathbf{y}(F)^{\mathrm{T}}\tilde{\mathbf{v}}=1-\delta$. This implies that $\|\mathbf{y}(F)-\tilde{\mathbf{v}}\|_2^2=2\delta$. Also, note that

$$\begin{aligned}\|\tilde{\mathbf{v}}-\mathbf{y}(F)\|^2_{\hat{\mathbf{B}}_F}&=\|\tilde{\mathbf{v}}-\alpha_1\boldsymbol{\xi}_1-\{\mathbf{y}(F)-\alpha_1\boldsymbol{\xi}_1\}\|^2_{\hat{\mathbf{B}}_F}\\ &=\|\tilde{\mathbf{v}}-\alpha_1\boldsymbol{\xi}_1\|^2_{\hat{\mathbf{B}}_F}+\|\mathbf{y}(F)-\alpha_1\boldsymbol{\xi}_1\|^2_{\hat{\mathbf{B}}_F}-2(\mathbf{y}(F)-\alpha_1\boldsymbol{\xi}_1)^{\mathrm{T}}\hat{\mathbf{B}}_F(\tilde{\mathbf{v}}-\alpha_1\boldsymbol{\xi}_1).\end{aligned}$$

Since $\mathbf{y}(F)-\alpha_1\boldsymbol{\xi}_1$ is orthogonal to $\tilde{\mathbf{v}}-\alpha_1\boldsymbol{\xi}_1$ under the normalization of $\hat{\mathbf{B}}_F$, we have

$$\sum_{j=2}^{k'}\alpha_j^2=\|\tilde{\mathbf{v}}-\alpha_1\boldsymbol{\xi}_1\|^2_{\hat{\mathbf{B}}_F}\leqslant\|\tilde{\mathbf{v}}-\mathbf{y}(F)\|^2_{\hat{\mathbf{B}}_F}\leqslant 2\,\lambda_{\max}(\hat{\mathbf{B}}_F)\delta, \tag{25}$$

in which the last inequality holds by an application of Hölder's inequality and the fact that $\|\mathbf{y}(F)-\tilde{\mathbf{v}}\|_2^2=2\delta$. Moreover, we have

$$\begin{gathered}\sum_{j=1}^{k'}\alpha_j^2=\|\tilde{\mathbf{v}}\|^2_{\hat{\mathbf{B}}_F}\geqslant\lambda_{\max}(\hat{\mathbf{B}}_F)/\hat{\kappa},\\ \alpha_1^2\geqslant\lambda_{\max}(\hat{\mathbf{B}}_F)/\hat{\kappa}-\sum_{j=2}^{k'}\alpha_j^2\geqslant 2\lambda_{\max}(\hat{\mathbf{B}}_F)/(3\hat{\kappa}),\end{gathered} \tag{26}$$

where the last inequality is obtained by result (25) and the assumption that $\delta\leqslant 1/(8c_{\mathrm{upper}}\kappa)$.

We also need a lower bound on $\|\mathbf{y}(F)\|_{\hat{\mathbf{B}}_F}$. By the triangle inequality, we have

$$\begin{aligned}\|\mathbf{y}(F)\|_{\hat{\mathbf{B}}_F}&\geqslant\|\tilde{\mathbf{v}}\|_{\hat{\mathbf{B}}_F}-\|\tilde{\mathbf{v}}-\mathbf{y}(F)\|_{\hat{\mathbf{B}}_F}\geqslant\sqrt{\left(\sum_{j=1}^{k'}\alpha_j^2\right)}-\surd\lambda_{\max}(\hat{\mathbf{B}}_F)\|\tilde{\mathbf{v}}-\mathbf{y}(F)\|_2\\ &\geqslant\frac{1}{2}\sqrt{\left(\sum_{j=1}^{k'}\alpha_j^2\right)}+\frac{1}{2}\sqrt{\left\{\frac{\lambda_{\max}(\hat{\mathbf{B}}_F)}{\hat{\kappa}}\right\}}-\surd\{2\lambda_{\max}(\hat{\mathbf{B}}_F)\delta\}\geqslant\frac{1}{2}\alpha_1,\end{aligned} \tag{27}$$

where the second inequality holds by the definition of $\|\tilde{\mathbf{v}}\|_{\hat{\mathbf{B}}_F}$ and an application of Hölder's inequality, the third inequality follows from expression (26), and the last inequality follows from the fact that

$\frac{1}{2}\surd\{\lambda_{\max}(\hat{\mathbf{B}}_F)/\hat{\kappa}\} \geqslant \surd\{2\lambda_{\max}(\hat{\mathbf{B}}_F)\delta\}$ under the assumption that $\delta \leqslant 1/(8c_{\text{upper}}\kappa)$.

C.1.1. Lower and upper bounds for $\{\hat{\lambda}_1(F)-\rho\}/\rho$

To obtain a lower bound for the quantity $\mathbf{y}(F)^{\mathrm{T}}\mathbf{v}'$, we need both the lower bound and the upper bound for the quantity $\{\hat{\lambda}_1(F)-\rho\}/\rho$. Recall that $\rho=\tilde{\mathbf{v}}^{\mathrm{T}}\hat{\mathbf{A}}_F\tilde{\mathbf{v}}/\tilde{\mathbf{v}}^{\mathrm{T}}\hat{\mathbf{B}}_F\tilde{\mathbf{v}}$. Using the fact that $\tilde{\mathbf{v}}^{\mathrm{T}}\hat{\mathbf{A}}_F\tilde{\mathbf{v}}=\Sigma_{j=1}^{k'}\alpha_j^2\hat{\lambda}_j(F)$, we obtain

$$\frac{\hat{\lambda}_1(F)-\rho}{\rho}=\frac{\sum_{j=1}^{k'}\{\hat{\lambda}_1(F)-\hat{\lambda}_j(F)\}\alpha_j^2}{\sum_{j=1}^{k'}\hat{\lambda}_j(F)\alpha_j^2}\leqslant\frac{\hat{\lambda}_1(F)\sum_{j=2}^{k'}\alpha_j^2}{\hat{\lambda}_1(F)\alpha_1^2}\leqslant\frac{2\lambda_{\max}(\hat{\mathbf{B}}_F)\delta}{\alpha_1^2}\leqslant 3\delta\hat{\kappa}, \tag{28}$$

where the second-to-last inequality holds by result (25) and the last inequality holds by expression (26). We now establish a lower bound for $\{\hat{\lambda}_1(F)-\rho\}/\rho$. First, we observe that

$$\delta\leqslant 2\delta-\delta^2=(1-\delta)^2+1-2(1-\delta)\mathbf{y}(F)^{\mathrm{T}}\tilde{\mathbf{v}}=\|\tilde{\mathbf{v}}-(1-\delta)\mathbf{y}(F)\|_2^2\leqslant\|\tilde{\mathbf{v}}-\alpha_1\boldsymbol{\xi}_1\|_2^2, \tag{29}$$

where the first equality follows from the fact that $\mathbf{y}(F)^{\mathrm{T}}\tilde{\mathbf{v}}=1-\delta$, and the second inequality holds by the fact that $(1-\delta)\mathbf{y}(F)$ is the scalar projection of $\mathbf{y}(F)$ onto the vector $\boldsymbol{\xi}_1$. Thus, we have

$$\begin{aligned}\frac{\hat{\lambda}_1(F)-\rho}{\rho}&=\frac{\sum_{j=1}^{k'}\{\hat{\lambda}_1(F)-\hat{\lambda}_j(F)\}\alpha_j^2}{\sum_{j=1}^{k'}\hat{\lambda}_j(F)\alpha_j^2}\geqslant\frac{\{\hat{\lambda}_1(F)-\hat{\lambda}_2(F)\}\sum_{j=2}^{k'}\alpha_j^2}{\hat{\lambda}_1(F)\alpha_1^2+\hat{\lambda}_2(F)\sum_{j=2}^{k'}\alpha_j^2}\\&=\frac{\{\hat{\lambda}_1(F)-\hat{\lambda}_2(F)\}\|\tilde{\mathbf{v}}-\alpha_1\boldsymbol{\xi}_1\|_{\hat{\mathbf{B}}_F}^2}{\hat{\lambda}_1(F)\alpha_1^2+\hat{\lambda}_2(F)\|\tilde{\mathbf{v}}-\alpha_1\boldsymbol{\xi}_1\|_{\hat{\mathbf{B}}_F}^2}\geqslant\frac{(1-\gamma)\{\lambda_{\max}(\hat{\mathbf{B}}_F)/\hat{\kappa}\}\|\tilde{\mathbf{v}}-\alpha_1\boldsymbol{\xi}_1\|_2^2}{\alpha_1^2+\gamma\{\lambda_{\max}(\hat{\mathbf{B}}_F)/\hat{\kappa}\}\|\tilde{\mathbf{v}}-\alpha_1\boldsymbol{\xi}_1\|_2^2}\\&\geqslant\frac{(1-\gamma)\,\lambda_{\max}(\hat{\mathbf{B}}_F)\delta}{\alpha_1^2\hat{\kappa}+\gamma\,\lambda_{\max}(\hat{\mathbf{B}}_F)\delta},\end{aligned} \tag{30}$$

where the second-to-last inequality holds by dividing the numerator and denominator by $\hat{\lambda}_1(F)$ and using the upper bound $\hat{\lambda}_2(F)/\hat{\lambda}_1(F)\leqslant\gamma$, and the last inequality holds by result (29).

C.1.2. Lower bound for $\|\mathbf{C}_F\tilde{\mathbf{v}}\|_2^{-1}$

In what follows, we first establish an upper bound for $\|\mathbf{C}_F\tilde{\mathbf{v}}\|_2^2$. By the definition that $\rho=\tilde{\mathbf{v}}^{\mathrm{T}}\hat{\mathbf{A}}_F\tilde{\mathbf{v}}/\tilde{\mathbf{v}}^{\mathrm{T}}\hat{\mathbf{B}}_F\tilde{\mathbf{v}}$, we have

$$\tilde{\mathbf{v}}^{\mathrm{T}}\hat{\mathbf{A}}_F\tilde{\mathbf{v}}-\rho\tilde{\mathbf{v}}^{\mathrm{T}}\hat{\mathbf{B}}_F\tilde{\mathbf{v}}=0.$$

Moreover, by the definition of $\tilde{\mathbf{v}}=\Sigma_{j=1}^{k'}\alpha_j\boldsymbol{\xi}_j$ and the fact that $\hat{\mathbf{A}}_F\boldsymbol{\xi}_j=\hat{\lambda}_j(F)\hat{\mathbf{B}}_F\boldsymbol{\xi}_j$, we have

$$\|(\hat{\mathbf{A}}_F-\rho\hat{\mathbf{B}}_F)\tilde{\mathbf{v}}\|_2^2=\left\|\sum_{j=1}^{k'}\alpha_j\hat{\mathbf{A}}_F\boldsymbol{\xi}_j-\rho\sum_{j=1}^{k'}\alpha_j\hat{\mathbf{B}}_F\boldsymbol{\xi}_j\right\|_2^2=\left\|\sum_{j=1}^{k'}\alpha_j\{\hat{\lambda}_j(F)-\rho\}\hat{\mathbf{B}}_F\boldsymbol{\xi}_j\right\|_2^2. \tag{31}$$

Thus, by equation (31) and the fact that $\tilde{\mathbf{v}}^{\mathrm{T}}\hat{\mathbf{A}}_F\tilde{\mathbf{v}}-\rho\tilde{\mathbf{v}}^{\mathrm{T}}\hat{\mathbf{B}}_F\tilde{\mathbf{v}}=0$, we obtain

$$\|\mathbf{C}_F\tilde{\mathbf{v}}\|_2^2=\left\|\left\{\mathbf{I}+\frac{\eta}{\rho}(\hat{\mathbf{A}}_F-\rho\hat{\mathbf{B}}_F)\right\}\tilde{\mathbf{v}}\right\|_2^2=1+\left\|\sum_{j=1}^{k'}\alpha_j\frac{\eta}{\rho}\{\hat{\lambda}_j(F)-\rho\}\hat{\mathbf{B}}_F\boldsymbol{\xi}_j\right\|_2^2. \tag{32}$$

It remains to establish an upper bound for the second term in equation (32). By the assumption that $\delta\leqslant 1/(3c_{\text{upper}}\kappa)(1/\gamma-1)$ and result (28), we have

$$\hat{\lambda}_2(F)\leqslant\rho\leqslant\hat{\lambda}_1(F).$$

Moreover, since $\|\tilde{\mathbf{v}}\|_2^2=1$, we have $\alpha_1^2\leqslant\lambda_{\max}(\hat{\mathbf{B}}_F)$. Thus,

$$\begin{aligned}\left\|\sum_{j=1}^{k'}\alpha_j(\eta/\rho)\{\hat\lambda_j(F)-\rho\}\hat{\mathbf{B}}_F\boldsymbol{\xi}_j\right\|_2^2 &\leqslant \alpha_1^2\{\hat\lambda_1(F)-\rho\}^2\lambda_{\max}(\hat{\mathbf{B}}_F)(\eta/\rho)^2+\lambda_{\max}(\hat{\mathbf{B}}_F)\sum_{j=2}^{k'}\alpha_j^2(\eta/\rho)^2\{\hat\lambda_j(F)-\rho\}^2\\ &\leqslant \lambda_{\max}^2(\hat{\mathbf{B}}_F)\eta^2(3\delta\hat\kappa)^2+\lambda_{\max}(\hat{\mathbf{B}}_F)\eta^2\{\hat\lambda_1(F)/\rho-1\}^2\sum_{j=2}^{k'}\alpha_j^2\\ &\leqslant \lambda_{\max}^2(\hat{\mathbf{B}}_F)\eta^2(3\delta\hat\kappa)^2+2\lambda_{\max}^2(\hat{\mathbf{B}}_F)\eta^2\delta(3\delta\hat\kappa)^2\\ &=9\,\lambda_{\max}^2(\hat{\mathbf{B}}_F)\eta^2\delta^2\hat\kappa^2+18\,\lambda_{\max}^2(\hat{\mathbf{B}}_F)\eta^2\delta^3\hat\kappa^2,\end{aligned}\tag{33}$$

where the second inequality is from result (28) and the third inequality follows from result (25). Substituting equation (33) into equation (32), we have

$$\begin{aligned}\|\mathbf{C}_F\tilde{\mathbf{v}}\|_2^2 &\leqslant 1+9\,\lambda_{\max}^2(\hat{\mathbf{B}}_F)\eta^2\delta^2\hat\kappa^2+18\,\lambda_{\max}^2(\hat{\mathbf{B}}_F)\eta^2\delta^3\hat\kappa^2\\ &\leqslant 1+12\,\lambda_{\max}^2(\hat{\mathbf{B}}_F)\eta^2\delta^2\hat\kappa^2,\end{aligned}\tag{34}$$

where the last inequality follows from the fact that $2\delta\leqslant\frac{1}{4}$, which holds by the assumption that $\delta\leqslant 1/(8c_{\text{upper}}\kappa)$. Meanwhile, note that the second term in the upper bound is less than 1 by the assumption $\delta\leqslant 1/(8c_{\text{upper}}\kappa)$ and $\eta c_{\text{upper}}\lambda_{\max}(\mathbf{B})<1$. Hence, by invoking inequality (34) and the fact that $1/\surd(1+y)\geqslant 1-y/2$ for $|y|<1$, we have

$$\|\mathbf{C}_F\tilde{\mathbf{v}}\|_2^{-1}\geqslant 1-6\,\lambda_{\max}^2(\hat{\mathbf{B}}_F)\eta^2\delta^2\hat\kappa^2.\tag{35}$$

C.1.3. Lower bound for $\mathbf{y}(F)^{\mathrm{T}}\mathbf{C}_F\tilde{\mathbf{v}}$

We have

$$\begin{aligned}\mathbf{y}(F)^{\mathrm{T}}\mathbf{C}_F\tilde{\mathbf{v}} &=\mathbf{y}(F)^{\mathrm{T}}\tilde{\mathbf{v}}+\frac{\eta}{\rho}\mathbf{y}(F)^{\mathrm{T}}(\hat{\mathbf{A}}_F-\rho\hat{\mathbf{B}}_F)\tilde{\mathbf{v}}\\ &=1-\delta+\frac{\eta}{\rho}\{\hat\lambda_1(F)-\rho\}\,\mathbf{y}(F)^{\mathrm{T}}\hat{\mathbf{B}}_F\tilde{\mathbf{v}}\\ &=1-\delta+\frac{\eta}{\rho}\{\hat\lambda_1(F)-\rho\}\alpha_1\frac{\boldsymbol{\xi}_1^{\mathrm{T}}\hat{\mathbf{B}}_F\boldsymbol{\xi}_1}{\|\boldsymbol{\xi}_1\|_2}\\ &=1-\delta+\eta\alpha_1\frac{\hat\lambda_1(F)-\rho}{\rho}\|\mathbf{y}(F)\|_{\hat{\mathbf{B}}_F}\\ &\geqslant 1-\delta+\frac{1}{2}\eta\alpha_1^2\frac{(1-\gamma)\,\lambda_{\max}(\hat{\mathbf{B}}_F)\delta}{\alpha_1^2\hat\kappa+\gamma\,\lambda_{\max}(\hat{\mathbf{B}}_F)\delta}\\ &\geqslant 1-\delta+\frac{1}{2}\eta\frac{\alpha_1^2(1-\gamma)\delta}{\hat\kappa+\gamma}\\ &\geqslant 1-\delta+\frac{1}{3}\eta\,\lambda_{\min}(\hat{\mathbf{B}}_F)\frac{(1-\gamma)\delta}{\hat\kappa+\gamma},\end{aligned}\tag{36}$$

where the first inequality follows from expressions (27) and (30), the second inequality uses the fact that $\alpha_1^2\leqslant\lambda_{\max}(\hat{\mathbf{B}}_F)$ and the last inequality follows from expression (26).

C.1.4. Combining the results

We now establish a lower bound on $\mathbf{y}(F)^{\mathrm{T}}\mathbf{v}'$. From expressions (35) and (36), we have

$$\begin{aligned}\mathbf{y}(F)^{\mathrm{T}}\mathbf{v}' &=\mathbf{y}(F)^{\mathrm{T}}\mathbf{C}_F\tilde{\mathbf{v}}\cdot\|\mathbf{C}_F\tilde{\mathbf{v}}\|_2^{-1}\\ &\geqslant\left\{1-\delta+\frac{1}{3}\eta\,\lambda_{\min}(\hat{\mathbf{B}}_F)\frac{(1-\gamma)\delta}{\hat\kappa+\gamma}\right\}\{1-6\,\lambda_{\max}^2(\hat{\mathbf{B}}_F)\eta^2\delta^2\hat\kappa^2\}\\ &\geqslant 1-\delta+\frac{1}{3}\eta\,\lambda_{\min}(\hat{\mathbf{B}}_F)\frac{(1-\gamma)\delta}{\hat\kappa+\gamma}-6\,\lambda_{\max}^2(\hat{\mathbf{B}}_F)\eta^2\delta^2\hat\kappa^2-2\hat\kappa^2\eta^3\lambda_{\max}^3(\hat{\mathbf{B}}_F)\delta^2\frac{(1-\gamma)\delta}{\hat\kappa+\gamma}\end{aligned}$$

$$\geqslant 1-\delta+\frac{1}{3}\eta\,\lambda_{\min}(\hat{\mathbf{B}}_F)\frac{(1-\gamma)\delta}{\hat{\kappa}+\gamma}-6.25\,\lambda^2_{\max}(\hat{\mathbf{B}}_F)\eta^2\delta^2\hat{\kappa}^2$$

$$\geqslant 1-\delta+\frac{1}{8}\eta\,\lambda_{\min}(\hat{\mathbf{B}}_F)\frac{(1-\gamma)\delta}{\hat{\kappa}+\gamma}, \tag{37}$$

in which the third inequality holds by the assumption that the step size η is sufficiently small such that $\eta\,\lambda_{\max}(\hat{\mathbf{B}}_F)<1$, and the last inequality holds under the condition that

$$\frac{1-\gamma}{\hat{\kappa}+\gamma}\geqslant 30\eta\,\lambda_{\max}(\hat{\mathbf{B}})\delta\hat{\kappa}^2,$$

which is implied by the following inequality under assumption 1:

$$\delta\leqslant\frac{1-\gamma}{30(1+c)c^2_{\text{upper}}\eta\,\lambda_{\max}(\mathbf{B})\kappa^2(c_{\text{upper}}\kappa+\gamma)}.$$

By assumption 1, we have

$$\mathbf{y}(F)^{\mathrm{T}}\mathbf{v}'\geqslant 1-\delta+\frac{1+c}{8}\eta\,\lambda_{\min}(\mathbf{B})\{1-\mathbf{y}(F)^{\mathrm{T}}\tilde{\mathbf{v}}\}\,\frac{1-\gamma}{c_{\text{upper}}\kappa+\gamma},$$

as desired.

C.2. Proof of lemma 5

The first part of lemma 5 on the following inequality follows directly from theorem 4.3 in Stewart (1979):

$$\frac{\|\mathbf{v}(F)-\mathbf{v}^*\|_2}{\|\mathbf{v}^*\|_2}\leqslant\frac{\delta(F)}{\Delta\hat{\lambda}\,\mathrm{cr}(\hat{\mathbf{A}}_F,\hat{\mathbf{B}}_F)}.$$

We now prove the second part of the lemma.

Setting $\mathbf{y}(F)=\mathbf{v}(F)/\|\mathbf{v}(F)\|_2$ and $\mathbf{y}^*=\mathbf{v}^*/\|\mathbf{v}^*\|_2$ such that $\|\mathbf{y}(F)\|_2=1$ and $\|\mathbf{y}^*\|_2=1$, we have

$$\begin{aligned}\|\mathbf{y}(F)-\mathbf{y}^*\|_2&\leqslant\left\|\frac{\mathbf{v}(F)}{\|\mathbf{v}(F)\|_2}-\frac{\mathbf{v}^*}{\|\mathbf{v}^*\|_2}\right\|_2\\&\leqslant\frac{1}{\|\mathbf{v}(F)\|_2\|\mathbf{v}^*\|_2}\|\mathbf{v}(F)\|\mathbf{v}^*\|_2-\mathbf{v}^*\|\mathbf{v}(F)\|_2\|_2\\&\leqslant\frac{2}{\|\mathbf{v}^*\|_2}\|\mathbf{v}(F)-\mathbf{v}^*\|_2\\&\leqslant 2\frac{\delta(F)}{\Delta\hat{\lambda}\,\mathrm{cr}(\hat{\mathbf{A}}_F,\hat{\mathbf{B}}_F)}\end{aligned}$$

where the third inequality holds by adding and subtracting $\mathbf{v}(F)\,\|\mathbf{v}(F)\|_2$. By definition, $\delta(F)\leqslant\epsilon(k')$ and $\Delta\hat{\lambda}\geqslant\Delta\lambda$. Moreover, by theorem 2.4 in Stewart (1979), $\mathrm{cr}(\hat{\mathbf{A}}_F,\hat{\mathbf{B}}_F)\geqslant\mathrm{cr}(k')-\epsilon(k')$. Thus, we obtain

$$\|\mathbf{y}(F)-\mathbf{y}^*\|_2\leqslant\frac{2}{\Delta\lambda\{\mathrm{cr}(k')-\epsilon(k')\}}\,\epsilon(k').$$

The other case for $\|\mathbf{y}(F)+\mathbf{y}^*\|_2$ can be proved similarly.

Appendix D: Proof of proposition 1

The proof of proposition 1 is an adaptation of the proof of theorem 4.1 in Gao *et al.* (2017) and the proof of theorem 1 in Tan *et al.* (2018), with some modifications to the curvature lemma to remove the structural assumptions on $\mathbf{A}$. Without loss of generality, we assume that $\mathbf{A}$ is full rank. For ease of notation, throughout the proof, we write $\mathbf{V}$, $\mathbf{\Lambda}$ and $\mathbf{P}$ to indicate $\mathbf{V}^*$, $\mathbf{\Lambda}^*$ and $\mathbf{P}^*$ respectively.

Let $\mathbf{V}=(\mathbf{V}_{\cdot K},\mathbf{V}_{\cdot K^c})\in\mathbb{R}^{d\times d}$, where $\mathbf{V}_{\cdot K}\in\mathbb{R}^{d\times K}$ are the K leading generalized eigenvectors of $(\mathbf{A},\mathbf{B})$ and

$\mathbf{V}_{\cdot K^c} \in \mathbb{R}^{d\times(d-K)}$ are the last $d-K$ generalized eigenvectors. Let $\mathbf{\Lambda} \in \mathbb{R}^{d\times d}$ be a diagonal matrix of the generalized eigenvalues. Let $\mathcal{S}_v$ be a set containing indices of non-zero rows of $\mathbf{V} \in \mathbb{R}^{d\times d}$, with cardinality $|\mathcal{S}_v| = s$. In other words, each generalized eigenvector has at most s non-zero elements. Let $\mathbf{P} = \mathbf{V}_{\cdot K}\mathbf{V}_{\cdot K}^{\mathrm{T}}$ and let $\mathcal{S}$ and $\mathcal{S}^c$ be the support of $\mathbf{P}$ and complementary set of $\mathcal{S}$ respectively. To facilitate the proof, we define some new notation:

$$\begin{aligned}\tilde{\mathbf{A}} &= \hat{\mathbf{B}}\mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}}\hat{\mathbf{B}},\\ \tilde{\mathbf{V}}_{\cdot K} &= \mathbf{V}_{\cdot K}(\mathbf{V}_{\cdot K}^{\mathrm{T}}\hat{\mathbf{B}}\mathbf{V}_{\cdot K})^{-1/2},\\ \tilde{\mathbf{P}} &= \tilde{\mathbf{V}}_{\cdot K}\tilde{\mathbf{V}}_{\cdot K}^{\mathrm{T}}.\end{aligned}$$

Let $\hat{\mathbf{P}}$ be a solution of problem (10) with tuning parameter K and ζ, and let $\mathbf{\Delta} = \hat{\mathbf{P}} - \tilde{\mathbf{P}}$. Finally, for two matrices $\mathbf{E}$ and $\mathbf{F}$, we write $\langle \mathbf{E}, \mathbf{F}\rangle = \mathrm{tr}(\mathbf{EF})$.

It can be shown that $\tilde{\mathbf{P}}$ satisfies both constraints in problem (10) and therefore is a feasible solution of problem (10). Since $\tilde{\mathbf{P}}$ is a feasible solution of problem (10) and $\hat{\mathbf{P}}$ is the optimal solution of problem (10), we have

$$-\langle \hat{\mathbf{A}}, \hat{\mathbf{P}}\rangle + \zeta\|\hat{\mathbf{P}}\|_{1,1} \leqslant -\langle \hat{\mathbf{A}}, \tilde{\mathbf{P}}\rangle + \zeta\|\tilde{\mathbf{P}}\|_{1,1}.$$

By picking $\zeta > 2\|\hat{\mathbf{A}} - \tilde{\mathbf{A}}\|_{\infty,\infty}$, the triangle inequality, rearranging the terms and using the fact that $\tilde{\mathbf{P}}$ and $\mathbf{P}$ share the same support, it can be shown that

$$-\langle \tilde{\mathbf{A}}, \mathbf{\Delta}\rangle \leqslant \frac{3\zeta}{2}\|\mathbf{\Delta}_{\mathcal{S}}\|_{1,1} - \frac{\zeta}{2}\|\mathbf{\Delta}_{\mathcal{S}^c}\|_{1,1}. \tag{38}$$

The main difference between our proof and that of Gao *et al.* (2017) and Tan *et al.* (2018) is in obtaining the lower bound for $-\langle \tilde{\mathbf{A}}, \mathbf{\Delta}\rangle$. By the definition of $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{P}}$, we obtain

$$\begin{aligned}-\langle \tilde{\mathbf{A}}, \mathbf{\Delta}\rangle &= \langle \hat{\mathbf{B}}\mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}}\hat{\mathbf{B}}, \tilde{\mathbf{P}} - \hat{\mathbf{P}}\rangle\\ &= \langle \hat{\mathbf{B}}^{1/2}\mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}}\hat{\mathbf{B}}^{1/2}, \hat{\mathbf{B}}^{1/2}(\tilde{\mathbf{P}} - \hat{\mathbf{P}})\hat{\mathbf{B}}^{1/2}\rangle\\ &= \langle \hat{\mathbf{B}}^{1/2}\mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}}\hat{\mathbf{B}}\tilde{\mathbf{P}}\hat{\mathbf{B}}^{1/2}, \mathbf{I} - \hat{\mathbf{B}}^{1/2}\hat{\mathbf{P}}\hat{\mathbf{B}}^{1/2}\rangle - \langle(\mathbf{I} - \hat{\mathbf{B}}^{1/2}\tilde{\mathbf{P}}\hat{\mathbf{B}}^{1/2})\hat{\mathbf{B}}^{1/2}\mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}}\hat{\mathbf{B}}^{1/2}, \hat{\mathbf{B}}^{1/2}\hat{\mathbf{P}}\hat{\mathbf{B}}^{1/2}\rangle\\ &= \mathrm{I} - \mathrm{II}.\end{aligned} \tag{39}$$

It suffices to obtain a lower bound for I and an upper bound for II.

D.1. Lower bound for I

We have

$$\begin{aligned}\mathrm{I} &= \mathrm{tr}\{\hat{\mathbf{B}}^{1/2}\mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}}\hat{\mathbf{B}}\mathbf{V}_{\cdot K}(\mathbf{V}_{\cdot K}^{\mathrm{T}}\hat{\mathbf{B}}\mathbf{V}_{\cdot K})^{-1}\mathbf{V}_{\cdot K}^{\mathrm{T}}\hat{\mathbf{B}}^{1/2}(\mathbf{I} - \hat{\mathbf{B}}^{1/2}\hat{\mathbf{P}}\mathbf{B}^{1/2})\}\\ &= \mathrm{tr}\{\hat{\mathbf{B}}_{\mathcal{S}_v}\mathbf{V}_{\mathcal{S}_v,K}(\mathbf{V}_{\cdot K}^{\mathrm{T}}\hat{\mathbf{B}}\mathbf{V}_{\cdot K})^{-1}\mathbf{V}_{\cdot K}^{\mathrm{T}}\hat{\mathbf{B}}^{1/2}(\mathbf{I} - \hat{\mathbf{B}}^{1/2}\hat{\mathbf{P}}\mathbf{B}^{1/2})\hat{\mathbf{B}}^{1/2}\mathbf{V}\mathbf{\Lambda}\mathbf{V}_{\mathcal{S}_v\cdot}^{\mathrm{T}}\}\\ &\geqslant \frac{\lambda_{\min}(\hat{\mathbf{B}}_{\mathcal{S}_v})}{\lambda_{\max}(\mathbf{B}_{\mathcal{S}_v})}\mathrm{tr}\{\hat{\mathbf{B}}^{1/2}\mathbf{V}\mathbf{\Lambda}\mathbf{V}_{\mathcal{S}_v\cdot}^{\mathrm{T}}\mathbf{B}_{\mathcal{S}_v}\mathbf{V}_{\mathcal{S}_v,K}(\mathbf{V}_{\cdot K}^{\mathrm{T}}\hat{\mathbf{B}}\mathbf{V}_{\cdot K})^{-1}\mathbf{V}_{\cdot K}^{\mathrm{T}}\hat{\mathbf{B}}^{1/2}(\mathbf{I} - \hat{\mathbf{B}}^{1/2}\hat{\mathbf{P}}\hat{\mathbf{B}}^{1/2})\}\\ &\geqslant \frac{\lambda_{\min}(\mathbf{B}_{\mathcal{S}_v}) - \rho(\mathbf{E}_{\mathbf{B}}, s)}{\lambda_{\max}(\mathbf{B}_{\mathcal{S}_v})}\mathrm{tr}\{\hat{\mathbf{B}}^{1/2}\mathbf{V}\mathbf{\Lambda}\mathbf{V}_{\mathcal{S}_v\cdot}^{\mathrm{T}}\mathbf{B}_{\mathcal{S}_v}\mathbf{V}_{\mathcal{S}_v,K}(\mathbf{V}_{\cdot K}^{\mathrm{T}}\hat{\mathbf{B}}\mathbf{V}_{\cdot K})^{-1}\mathbf{V}_{\cdot K}^{\mathrm{T}}\hat{\mathbf{B}}^{1/2}(\mathbf{I} - \hat{\mathbf{B}}^{1/2}\hat{\mathbf{P}}\hat{\mathbf{B}}^{1/2})\}\\ &\geqslant \frac{1-c}{\kappa(\mathbf{B})}\mathrm{tr}\{\hat{\mathbf{B}}^{1/2}\mathbf{V}\mathbf{\Lambda}\mathbf{V}_{\mathcal{S}_v\cdot}^{\mathrm{T}}\mathbf{B}_{\mathcal{S}_v}\mathbf{V}_{\mathcal{S}_v,K}(\mathbf{V}_{\cdot K}^{\mathrm{T}}\hat{\mathbf{B}}\mathbf{V}_{\cdot K})^{-1}\mathbf{V}_{\cdot K}^{\mathrm{T}}\hat{\mathbf{B}}^{1/2}(\mathbf{I} - \hat{\mathbf{B}}^{1/2}\hat{\mathbf{P}}\hat{\mathbf{B}}^{1/2})\},\end{aligned} \tag{40}$$

where the second inequality holds by Weyl's inequality, i.e. $\lambda_{\min}(\mathbf{B}_{\mathcal{S}_v}) \leqslant \lambda_{\min}(\hat{\mathbf{B}}_{\mathcal{S}_v}) + \rho(\mathbf{E}_{\mathbf{B}}, s)$, and the last inequality follows from assumption 1. Note that

$$\begin{aligned}&\mathrm{tr}\{\hat{\mathbf{B}}^{1/2}\mathbf{V}\mathbf{\Lambda}\mathbf{V}_{\mathcal{S}_v\cdot}^{\mathrm{T}}\mathbf{B}_{\mathcal{S}_v}\mathbf{V}_{\mathcal{S}_v,K}(\mathbf{V}_{\cdot K}^{\mathrm{T}}\hat{\mathbf{B}}\mathbf{V}_{\cdot K})^{-1}\mathbf{V}_{\cdot K}^{\mathrm{T}}\hat{\mathbf{B}}^{1/2}(\mathbf{I} - \hat{\mathbf{B}}^{1/2}\hat{\mathbf{P}}\mathbf{B}^{1/2})\}\\ &\quad = \mathrm{tr}\left\{\hat{\mathbf{B}}^{1/2}(\mathbf{V}_{\cdot K}, \mathbf{V}_{\cdot K^c})\begin{pmatrix}\mathbf{\Lambda}_K & \mathbf{0}\\ \mathbf{0} & \mathbf{\Lambda}_{K^c}\end{pmatrix}\begin{pmatrix}\mathbf{I}_K\\ \mathbf{0}\end{pmatrix}(\mathbf{V}_{\cdot K}^{\mathrm{T}}\hat{\mathbf{B}}\mathbf{V}_{\cdot K})^{-1}\mathbf{V}_{\cdot K}^{\mathrm{T}}\hat{\mathbf{B}}^{1/2}(\mathbf{I} - \hat{\mathbf{B}}^{1/2}\hat{\mathbf{P}}\hat{\mathbf{B}}^{1/2})\right\}\\ &\quad = \mathrm{tr}\{\hat{\mathbf{B}}^{1/2}\mathbf{V}_{\cdot K}\mathbf{\Lambda}_K(\mathbf{V}_{\cdot K}^{\mathrm{T}}\hat{\mathbf{B}}\mathbf{V}_{\cdot K})^{-1}\mathbf{V}_{\cdot K}^{\mathrm{T}}\hat{\mathbf{B}}^{1/2}(\mathbf{I} - \hat{\mathbf{B}}^{1/2}\hat{\mathbf{P}}\hat{\mathbf{B}}^{1/2})\}.\end{aligned}$$

Substituting this into inequality (40), we obtain

$$\mathrm{I} \geqslant \frac{(1-c)\lambda_K}{\kappa(\mathbf{B})} \langle \hat{\mathbf{B}}^{1/2}\tilde{\mathbf{P}}\hat{\mathbf{B}}^{1/2}, \mathbf{I}-\hat{\mathbf{B}}^{1/2}\hat{\mathbf{P}}\hat{\mathbf{B}}^{1/2}\rangle. \tag{41}$$

D.2. Upper bound for II

Observe that

$$\begin{aligned}(\mathbf{I}-\hat{\mathbf{B}}^{1/2}\tilde{\mathbf{P}}\hat{\mathbf{B}}^{1/2})\hat{\mathbf{B}}^{1/2}\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}}\hat{\mathbf{B}}^{1/2} &= \hat{\mathbf{B}}^{1/2}\mathbf{V}_{\cdot K}\boldsymbol{\Lambda}_K\mathbf{V}_{\cdot K}^{\mathrm{T}}\hat{\mathbf{B}}^{1/2} + \hat{\mathbf{B}}^{1/2}\mathbf{V}_{\cdot K^{\mathrm{c}}}\boldsymbol{\Lambda}_{K^{\mathrm{c}}}\mathbf{V}_{\cdot K^{\mathrm{c}}}^{\mathrm{T}}\hat{\mathbf{B}}^{1/2} - \hat{\mathbf{B}}^{1/2}\mathbf{V}_{\cdot K}\boldsymbol{\Lambda}_K\mathbf{V}_{\cdot K}^{\mathrm{T}}\hat{\mathbf{B}}^{1/2}\\ &\quad - \hat{\mathbf{B}}^{1/2}\mathbf{V}_{\cdot K}(\mathbf{V}_{\cdot K}^{\mathrm{T}}\hat{\mathbf{B}}\mathbf{V}_{\cdot K})^{-1}\mathbf{V}_{\cdot K}^{\mathrm{T}}\hat{\mathbf{B}}\mathbf{V}_{\cdot K^{\mathrm{c}}}\boldsymbol{\Lambda}_{K^{\mathrm{c}}}\mathbf{V}_{\cdot K^{\mathrm{c}}}^{\mathrm{T}}\hat{\mathbf{B}}^{1/2}\\ &= (\mathbf{I}-\hat{\mathbf{B}}^{1/2}\tilde{\mathbf{P}}\hat{\mathbf{B}}^{1/2})\hat{\mathbf{B}}^{1/2}\mathbf{V}_{\cdot K^{\mathrm{c}}}\boldsymbol{\Lambda}_{K^{\mathrm{c}}}\mathbf{V}_{\cdot K^{\mathrm{c}}}^{\mathrm{T}}\hat{\mathbf{B}}^{1/2},\end{aligned}$$

where the last equality holds since the first equality depends only on $\boldsymbol{\Lambda}_{K^{\mathrm{c}}}$. Thus, we have

$$\begin{aligned}\mathrm{II} &= \langle(\mathbf{I}-\hat{\mathbf{B}}^{1/2}\tilde{\mathbf{P}}\hat{\mathbf{B}}^{1/2})\hat{\mathbf{B}}^{1/2}\mathbf{V}_{\cdot K^{\mathrm{c}}}\boldsymbol{\Lambda}_{K^{\mathrm{c}}}\mathbf{V}_{\cdot K^{\mathrm{c}}}^{\mathrm{T}}\hat{\mathbf{B}}^{1/2}, \hat{\mathbf{B}}^{1/2}\hat{\mathbf{P}}\hat{\mathbf{B}}^{1/2}\rangle\\ &\leqslant \lambda_{K+1}\langle(\mathbf{I}-\hat{\mathbf{B}}^{1/2}\tilde{\mathbf{P}}\hat{\mathbf{B}}^{1/2})\hat{\mathbf{B}}^{1/2}\mathbf{V}_{\cdot K^{\mathrm{c}}}\mathbf{V}_{\cdot K^{\mathrm{c}}}^{\mathrm{T}}\hat{\mathbf{B}}^{1/2}, \hat{\mathbf{B}}^{1/2}\hat{\mathbf{P}}\hat{\mathbf{B}}^{1/2}\rangle\\ &\leqslant \lambda_{K+1}\|\mathbf{V}_{\cdot K^{\mathrm{c}}}^{\mathrm{T}}\hat{\mathbf{B}}\mathbf{V}_{\cdot K^{\mathrm{c}}}\|_2\langle\mathbf{I}-\hat{\mathbf{B}}^{1/2}\tilde{\mathbf{P}}\hat{\mathbf{B}}^{1/2}, \hat{\mathbf{B}}^{1/2}\hat{\mathbf{P}}\hat{\mathbf{B}}^{1/2}\rangle\\ &\leqslant \lambda_{K+1}\{1+\|\mathbf{V}_{\cdot K^{\mathrm{c}}}^{\mathrm{T}}(\hat{\mathbf{B}}-\mathbf{B})\mathbf{V}_{\cdot K^{\mathrm{c}}}\|_2\}\langle\mathbf{I}-\hat{\mathbf{B}}^{1/2}\tilde{\mathbf{P}}\hat{\mathbf{B}}^{1/2}, \hat{\mathbf{B}}^{1/2}\hat{\mathbf{P}}\hat{\mathbf{B}}^{1/2}\rangle,\end{aligned} \tag{42}$$

where the last inequality holds by adding and subtracting $\mathbf{V}_{\cdot K^{\mathrm{c}}}^{\mathrm{T}}\mathbf{B}\mathbf{V}_{\cdot K^{\mathrm{c}}}$ and the triangle inequality. Since only s rows of $\mathbf{V}_{\cdot K^{\mathrm{c}}}$ are non-zero, by Holder's inequality, we obtain

$$\begin{aligned}\|\mathbf{V}_{\cdot K^{\mathrm{c}}}^{\mathrm{T}}(\hat{\mathbf{B}}-\mathbf{B})\mathbf{V}_{\cdot K^{\mathrm{c}}}\|_2 &\leqslant \|\mathbf{B}^{-1/2}\mathbf{B}^{1/2}\mathbf{V}_{\cdot K^{\mathrm{c}}}\|_2^2\,\rho(\mathbf{E}_{\mathbf{B}}, s)\\ &\leqslant c\lambda_{\min}(\mathbf{B})\|\mathbf{B}^{-1/2}\|_2^2\\ &\leqslant c,\end{aligned}$$

where the second inequality holds under assumption 1. Substituting this into inequality (42), we obtain

$$\mathrm{II} \leqslant c\lambda_{K+1}\langle\mathbf{I}-\hat{\mathbf{B}}^{1/2}\tilde{\mathbf{P}}\hat{\mathbf{B}}^{1/2}, \hat{\mathbf{B}}^{1/2}\hat{\mathbf{P}}\hat{\mathbf{B}}^{1/2}\rangle. \tag{43}$$

By definition, $\mathrm{tr}(\hat{\mathbf{B}}^{1/2}\hat{\mathbf{P}}\hat{\mathbf{B}}^{1/2}) = \mathrm{tr}(\hat{\mathbf{B}}^{1/2}\tilde{\mathbf{P}}\hat{\mathbf{B}}^{1/2}) = K$. Substituting inequalities (41) and (43) into inequality (39), we obtain

$$\begin{aligned}-\langle\tilde{\mathbf{A}}, \boldsymbol{\Delta}\rangle &\geqslant \left\{\frac{(1-c)\lambda_K}{\kappa(\mathbf{B})} - c\lambda_{K+1}\right\}(K - \langle\hat{\mathbf{B}}^{1/2}\hat{\mathbf{P}}\hat{\mathbf{B}}^{1/2}, \hat{\mathbf{B}}^{1/2}\tilde{\mathbf{P}}\hat{\mathbf{B}}^{1/2}\rangle)\\ &\geqslant \frac{1}{2}\left\{\frac{(1-c)\lambda_K}{\kappa(\mathbf{B})} - c\lambda_{K+1}\right\}\|\hat{\mathbf{B}}^{1/2}\boldsymbol{\Delta}\hat{\mathbf{B}}^{1/2}\|_{\mathrm{F}}^2.\end{aligned} \tag{44}$$

The rest of the proof follows from the proof of theorem 1 in Tan *et al.* (2018) or the proof of theorem 4.1 in Gao *et al.* (2017). We hereby provide a proof sketch and refer the reader to Tan *et al.* (2018) for the details. For notational convenience, let $\delta_{\mathrm{gap}} = (1-c)\lambda_K/\kappa(\mathbf{B}) - c\lambda_{K+1}$. Combining inequalities (38) and (44), we have

$$\|\hat{\mathbf{B}}^{1/2}\boldsymbol{\Delta}\hat{\mathbf{B}}^{1/2}\|_{\mathrm{F}}^2 \leqslant \frac{3\zeta}{\delta_{\mathrm{gap}}}\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{1,1}. \tag{45}$$

Moreover, $-\langle\tilde{\mathbf{A}}, \boldsymbol{\Delta}\rangle \geqslant 0$ implies that $\|\boldsymbol{\Delta}_{\mathcal{S}^{\mathrm{c}}}\|_{1,1} \leqslant 3\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{1,1}$.

Similarly to Tan *et al.* (2018), we partition the set $\mathcal{S}^{\mathrm{c}}$ into J sets such that $\mathcal{S}_1^{\mathrm{c}}$ is the index set of the largest l entries in absolute values of $\boldsymbol{\Delta}$, $\mathcal{S}_2^{\mathrm{c}}$ is the index set of the second largest l entries of $\boldsymbol{\Delta}$, and so forth, with $|\mathcal{S}_J^{\mathrm{c}}| \leqslant l$. By lemma S4 of Tan *et al.* (2018) and the fact that $\|\boldsymbol{\Delta}_{\mathcal{S}^{\mathrm{c}}}\|_{1,1} \leqslant 3\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{1,1}$, we obtain $\Sigma_{j=2}^J\|\boldsymbol{\Delta}_{\mathcal{S}_j^{\mathrm{c}}}\|_{\mathrm{F}} \leqslant 3sl^{-1/2}\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{\mathrm{F}}$. Under assumption 1, picking $l = c_1 s^2$, it can be shown that

$$\begin{aligned}\|\hat{\mathbf{B}}^{1/2}\boldsymbol{\Delta}\hat{\mathbf{B}}^{1/2}\|_{\mathrm{F}} &\geqslant \|\hat{\mathbf{B}}^{1/2}\boldsymbol{\Delta}_{\mathcal{S}\cup\mathcal{S}_1^{\mathrm{c}}}\hat{\mathbf{B}}^{1/2}\|_{\mathrm{F}} - \sum_{j=2}^{J}\|\hat{\mathbf{B}}^{1/2}\boldsymbol{\Delta}_{\mathcal{S}_j^{\mathrm{c}}}\hat{\mathbf{B}}^{1/2}\|_{\mathrm{F}}\\ &\geqslant \left\{(1-c)\lambda_{\min}(\mathbf{B}) - \frac{3(1+c)\lambda_{\max}(\mathbf{B})}{c_1}\right\}\|\boldsymbol{\Delta}_{\mathcal{S}\cup\mathcal{S}_1^{\mathrm{c}}}\|_{\mathrm{F}}\\ &\geqslant C\|\boldsymbol{\Delta}_{\mathcal{S}\cup\mathcal{S}_1^{\mathrm{c}}}\|_{\mathrm{F}},\end{aligned} \tag{46}$$

where C is a generic constant, and the last inequality holds by picking c_1 to be sufficiently large.

Combining inequalities (45) and (46),

$$\|\boldsymbol{\Delta}_{\mathcal{S}\cup\mathcal{S}_1^c}\|_F \leqslant C\left(\frac{\zeta}{\delta_{\text{gap}}}\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{1,1}\right)^{1/2} \leqslant C\left(\frac{\zeta s}{\delta_{\text{gap}}}\|\boldsymbol{\Delta}_{\mathcal{S}\cup\mathcal{S}_1^c}\|_F\right)^{1/2}.$$

By squaring both sides, we obtain $\|\boldsymbol{\Delta}_{\mathcal{S}\cup\mathcal{S}_1^c}\|_F \leqslant C\zeta s/\delta_{\text{gap}}$. By the triangle inequality,

$$\begin{aligned}\|\boldsymbol{\Delta}\|_F &\leqslant \|\boldsymbol{\Delta}_{\mathcal{S}\cup\mathcal{S}_1^c}\|_F + \|\boldsymbol{\Delta}_{(\mathcal{S}\cup\mathcal{S}_1^c)^c}\|_F \\ &\leqslant \|\boldsymbol{\Delta}_{\mathcal{S}\cup\mathcal{S}_1^c}\|_F + \sum_{j=2}^{J}\|\boldsymbol{\Delta}_{\mathcal{S}_j^c}\|_F \\ &\leqslant (1+3/c_1)\|\boldsymbol{\Delta}_{\mathcal{S}\cup\mathcal{S}_1^c}\|_F \\ &\leqslant C\frac{\zeta s}{\delta_{\text{gap}}}, \end{aligned} \tag{47}$$

where the second inequality holds by lemma S4 of Tan *et al.* (2018) and the third inequality holds by picking $l=c_1s^2$. Finally, by the triangle inequality and lemma S1 of Tan *et al.* (2018), we obtain

$$\|\hat{\mathbf{P}}-\mathbf{P}\|_F \leqslant \|\boldsymbol{\Delta}\|_F + \|\mathbf{P}-\tilde{\mathbf{P}}\|_F \leqslant C\left(\frac{\zeta s}{\delta_{\text{gap}}} + K\|\hat{\mathbf{B}}_{\mathcal{S}_v}-\mathbf{B}_{\mathcal{S}_v}\|_2\right).$$

References

d'Aspremont, A., Bach, F. and Ghaoui, L. E. (2008) Optimal solutions for sparse principal component analysis. *J. Mach. Learn. Res.*, **9**, 1269–1294.

d'Aspremont, A., El Ghaoui, L., Jordan, M. I. and Lanckriet, G. R. (2007) A direct formulation for sparse PCA using semidefinite programming. *SIAM Rev.*, **49**, 434–448.

Birnbaum, A., Johnstone, I. M., Nadler, B. and Paul, D. (2013) Minimax bounds for sparse PCA with noisy high-dimensional data. *Ann. Statist.*, **41**, 1055–1084.

Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2010) Distributed optimization and statistical learning via the ADMM. *Foundns Trends Mach. Learn.*, **3**, 1–122.

Cai, T. T., Ma, Z. and Wu, Y. (2013) Sparse PCA: optimal rates and adaptive estimation. *Ann. Statist.*, **41**, 3074–3110.

Chen, M., Gao, C., Ren, Z. and Zhou, H. H. (2013) Sparse CCA via precision adjusted iterative thresholding. *Preprint arXiv:1311.6186*. Department of Medicine, University of Chicago, Chicago.

Chen, X., Zou, C. and Cook, R. D. (2010) Coordinate-independent sparse sufficient dimension reduction and variable selection. *Ann. Statist.*, **38**, 3696–3723.

Clemmensen, L., Hastie, T., Witten, D. and Ersbøll, B. (2012) Sparse discriminant analysis. *Technometrics*, **53**, 406–413.

Cook, R. D. (2000) SAVE: a method for dimension reduction and graphics in regression. *Communs Statist. Theory Meth.*, **29**, 2109–2121.

Cook, R. D. (2007) Dimension reduction in regression. *Statist. Sci.*, **22**, 1–26.

Cook, R. D. and Forzani, L. (2008) Principal fitted components for dimension reduction in regression. *Statist. Sci.*, **23**, 485–501.

Cook, R. D. and Lee, H. (1999) Dimension reduction in binary response regression. *J. Am. Statist. Ass.*, **94**, 1187–1200.

Eckstein, J. (2012) Augmented Lagrangian and alternating direction methods for convex optimization: a tutorial and some illustrative computational results. *Research Report 32*. Rutgers Center for Operations Research, Piscataway.

Fan, J. and Fan, Y. (2008) High dimensional classification using features annealed independence rules. *Ann. Statist.*, **36**, 2605–2637.

Fan, J., Ke, Z. T., Liu, H. and Xia, L. (2015) QUADRO: a supervised dimension reduction method via Rayleigh quotient optimization. *Ann. Statist.*, **43**, 14–98.

Gao, C., Ma, Z., Ren, Z. and Zhou, H. H. (2015) Minimax estimation in sparse canonical correlation analysis. *Ann. Statist.*, **43**, 2168–2197.

Gao, C., Ma, Z. and Zhou, H. H. (2017) Sparse CCA: adaptive estimation and computational barriers. *Ann. Statist.*, **45**, 2074–2101.

Gaynanova, I. and Kolar, M. (2015) Optimal variable selection in multi-group sparse discriminant analysis. *Electron. J. Statist.*, **9**, 2007–2034.

Ge, R., Jin, C., Kakade, S. M., Netrapalli, P. and Sidford, A. (2016) Efficient algorithms for large-scale generalized eigenvector computation and canonical correlation analysis. In *Proc. 33rd Int. Conf. Machine Learning, New York*.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Golub, G. H. and Van Loan, C. F. (2012) *Matrix Computations*, vol. 3. Baltimore: Johns Hopkins University Press.

Gu, Q., Wang, Z. and Liu, H. (2014) Sparse PCA with oracle property. In *Proc. Conf. Advances in Neural Information Processing Systems, Montreal*.

Guo, Y., Hastie, T. and Tibshirani, R. (2007) Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, **8**, 86–100.

Kolar, M. and Liu, H. (2015) Optimal feature selection in high-dimensional discriminant analysis. *IEEE Trans. Inform. Theory*, **61**, 1063–1083.

Leng, C. (2008) Sparse optimal scoring for multiclass cancer diagnosis and biomarker detection using microarray data. *Computnl Biol. Chem.*, **32**, 417–425.

Li, K.-C. (1991) Sliced inverse regression for dimension reduction. *J. Am. Statist. Ass.*, **86**, 316–327.

Li, L. (2007) Sparse sufficient dimension reduction. *Biometrika*, **94**, 603–613.

Li, L. and Nachtsheim, C. J. (2006) Sparse sliced inverse regression. *Technometrics*, **48**, 503–510.

Li, L. and Yin, X. (2008) Sliced inverse regression with regularizations. *Biometrics*, **64**, 124–131.

Ma, Z. (2013) Sparse principal component analysis and iterative thresholding. *Ann. Statist.*, **41**, 772–801.

Ma, Z. and Li, X. (2016) Subspace perspective on canonical correlation analysis: dimension reduction and minimax rates. *Preprint arXiv:1605.03662*. Wharton School, University of Pennsylvania, Philadelphia.

Ma, Y. and Zhu, L. (2013) A review on dimension reduction. *Int. Statist. Rev.*, **81**, 134–150.

Mai, Q., Yang, Y. and Zou, H. (2015) Multiclass sparse discriminant analysis. *Statist. Sin.*, to be published.

Mai, Q., Zou, H. and Yuan, M. (2012) A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, **99**, 29–42.

Moghaddam, B., Weiss, Y. and Avidan, S. (2006a) Generalized spectral bounds for sparse LDA. In *Proc. 23rd Int. Conf. Machine Learning, Pittsburgh*.

Moghaddam, B., Weiss, Y. and Avidan, S. (2006b) Spectral bounds for sparse PCA: exact and greedy algorithms. In *Advances in Neural Information Processing Systems, Vancouver*.

Petersen, A., Witten, D. and Simon, N. (2016) Fused lasso additive model. *J. Computnl Graph. Statist.*, **25**, 1005–1025.

Spira, A., Beane, J. E., Shah, V., Steiling, K., Liu, G., Schembri, F., Gilman, S., Dumas, Y.-M., Calner, P., Sebastiani, P., Sridhar, S., Beamis, J., Lamb, C., Anderson, T., Gerry, N., Keane, J., Lenburg, M. E. and Brody, J. S. (2007) Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nat. Med.*, **13**, 361–366.

Stewart, G. (1979) Pertubation bounds for the definite generalized eigenvalue problem. *Lin. Alg. Appl.*, **23**, 69–85.

Stewart, G. and Sun, J. (1990) *Matrix Perturbation Theory*. New York: Elsevier.

Tan, K. M., Wang, Z., Zhang, T., Liu, H. and Cook, R. D. (2018) A convex formulation for high-dimensional sparse sliced inverse regression. *Biometrika*, to be published.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2003) Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statist. Sci.*, **18**, 104–117.

Vu, V. Q., Cho, J., Lei, J. and Rohe, K. (2013) Fantope projection and selection: a near-optimal convex relaxation of sparse PCA. In *Proc. Conf. Advances in Neural Information Processing Systems, Lake Tahoe*.

Vu, V. Q. and Lei, J. (2013) Minimax sparse principal subspace estimation in high dimensions. *Ann. Statist.*, **41**, 2905–2947.

Wang, Z., Han, F. and Liu, H. (2013) Sparse principal component analysis for high dimensional multivariate time series. In *Proc. 16th Int. Conf. Artificial Intelligence and Statistics, Scotsdale*.

Wang, Z., Lu, H. and Liu, H. (2014) Tighten after relax: minimax-optimal sparse PCA in polynomial time. In *Proc. Conf. Advances in Neural Information Processing Systems, Montreal*.

Witten, D. M., Tibshirani, R. and Hastie, T. (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, **10**, 515–534.

Yin, X. and Hilafu, H. (2015) Sequential sufficient dimension reduction for large *p*, small *n* problems. *J. R. Statist. Soc.* B, **77**, 879–892.

Yu, Y., Wang, T. and Samworth, R. J. (2014) A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, **102**, 315–323.

Yuan, X.-T. and Zhang, T. (2013) Truncated power method for sparse eigenvalue problems. *J. Mach. Learn. Res.*, **14**, 899–925.

Zhu, L., Miao, B. and Peng, H. (2006) On sliced inverse regression with high-dimensional covariates. *J. Am. Statist. Ass.*, **101**, 630–643.

Zou, H., Hastie, T. and Tibshirani, R. (2006) Sparse principal component analysis. *J. Computnl Graph. Statist.*, **15**, 265–286.