Sparse principal component analysis via random projections

Milana Gataric, Tengyao Wang and Richard J. Samworth

Statistical Laboratory, University of Cambridge

{m.gataric,t.wang,r.samworth}@statslab.cam.ac.uk

January 16, 2018

Abstract

We introduce a new method for sparse principal component analysis, based on the aggregation of eigenvector information from carefully-selected random projections of the sample covariance matrix. Unlike most alternative approaches, our algorithm is non-iterative, so is not vulnerable to a bad choice of initialization. Our theory provides great detail on the statistical and computational trade-off in our procedure, revealing a subtle interplay between the effective sample size and the number of random projections that are required to achieve the minimax optimal rate. Numerical studies provide further insight into the procedure and confirm its highly competitive finite-sample performance.

1 Introduction

Principal component analysis (PCA) is one of the most widely-used techniques for dimensionality reduction in Statistics, Image Processing and many other fields. The aim is to project the data along directions that explain the greatest proportion of the variance in the population. In the simplest setting where we seek a single, univariate projection of our data, we may estimate this optimal direction by computing the leading eigenvector of the sample covariance matrix.

Despite its successes and enormous popularity, it has been well-known for a decade or more that PCA breaks down as soon as the dimensionality p of the data is of the same order as the sample size n. More precisely, suppose that $X_1, \ldots, X_n \overset{\mathrm{iid}}{\sim} N_p(0, \Sigma)$ are observations from a Gaussian distribution with a spiked covariance matrix $\Sigma = I_p + v_1 v_1^\top$ whose leading eigenvector is $v_1 \in \mathcal{S}^{p-1} := \{v \in \mathbb{R}^p : \|v\| = 1\}$, and let $\hat{v}_1$ denote the leading unit-length eigenvector of the sample covariance matrix $\hat{\Sigma} := n^{-1} \sum_{i=1}^n X_i X_i^\top$. Then Johnstone and Lu

(2009) and Paul (2007) showed that $\hat{v}_1$ is a consistent estimator of v_1, i.e. $|\hat{v}_1^\top v_1| \xrightarrow{p} 1$, if and only if $p = p_n$ satisfies $p/n \to 0$ as $n \to \infty$. It is also worth noting that the principal component v_1 may be a linear combination of all elements of the canonical basis in $\mathbb{R}^p$, which can often make it difficult to interpret the estimated projected directions (Jolliffe, Trendafilov and Uddin, 2003).

To remedy this situation, and to provide additional interpretability to the principal components in high-dimensional settings, Jolliffe, Trendafilov and Uddin (2003) and Zou, Hastie and Tibshirani (2006) proposed Sparse Principal Component Analysis (SPCA). Here it is assumed that the leading population eigenvectors belong to the k-sparse unit ball

$$\mathcal{B}_0^{p-1}(k) := \left\{ v = (v^{(1)}, \ldots, v^{(p)})^\top \in \mathcal{S}^{p-1} : \sum_{j=1}^{p} \mathbb{1}_{\{v^{(j)} \neq 0\}} \leq k \right\}$$

for some $k \in \{1, \ldots, p\}$. In addition to the easier interpretability, a great deal of research effort has shown that such an assumption facilitates improved estimation performance (e.g. Johnstone and Lu, 2009; Paul and Johnstone, 2012; Vu and Lei, 2013; Cai, Ma and Wu, 2013; Ma, 2013; Wang, Berthet and Samworth, 2016a). To give a flavor of these results, let $\mathcal{V}_n$ denote the set of all estimators of v_1, i.e. the class of Borel measurable functions from $\mathbb{R}^{n\times p}$ to $\mathcal{S}^{p-1}$. Vu and Lei (2013) introduce a class $\mathcal{Q}$ of sub-Gaussian distributions whose first principal component v_1 belongs to $\mathcal{B}_0^{p-1}(k)$ and show that

$$\inf_{\tilde{v}_1 \in \mathcal{V}_n} \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q \{1 - (\tilde{v}_1^\top v_1)^2\} \asymp \frac{k \log p}{n}. \tag{1}$$

Here, $a_n \asymp b_n$ means $0 < \liminf_{n\to\infty} |a_n/b_n| \leq \limsup_{n\to\infty} |a_n/b_n| < \infty$. Thus, consistent estimation is possible in this framework provided only that $k = k_n$ and $p = p_n$ satisfy $(k \log p)/n \to 0$. Vu and Lei (2013) show further that this estimation rate is achieved by the natural estimator

$$\hat{v}_1 \in \operatorname*{argmax}_{v \in \mathcal{B}_0^{p-1}(k)} v^\top \hat{\Sigma} v. \tag{2}$$

However, results such as (1) do not complete the story of SPCA. Indeed, computing the estimator defined in (2) turns out to be an NP-hard problem (e.g. Tillmann and Pfetsch, 2014): the naive approach would require searching through all $\binom{p}{k}$ of the $k \times k$ symmetric submatrices of $\hat{\Sigma}$, which takes exponential time in k. Therefore, in parallel to the theoretical developments described above, numerous alternative algorithms for SPCA have been proposed in recent years. For instance, several papers have introduced techniques based on solving the non-convex optimization problem in (2) by invoking an ℓ_1-penalty (e.g. Jolliffe, Trendafilov and Uddin, 2003; Zou, Hastie and Tibshirani, 2006; Shen and Huang, 2008; Witten, Tibshirani and Hastie, 2009). Typically, these methods are fast, but lack theoretical performance guarantees. On the other hand, d'Aspremont et al. (2007) propose to solve the optimization problem in (2) via semidefinite relaxation. This approach was analyzed in the work of Amini and Wainwright (2009) and Wang, Berthet and Samworth (2016a), and has been proved to achieve the minimax rate of convergence under certain assumptions on the underlying distribution and asymptotic regime, but the algorithm is slow compared to other approaches. In a separate, recent development, it is now understood that, conditional on a Planted Clique hypothesis from theoretical computer science, there is an asymptotic regime in which no randomized polynomial time algorithm can attain the minimax optimal rate (Wang, Berthet and Samworth, 2016a). Various fast, iterative algorithms were introduced by Johnstone and Lu (2009), Paul and Johnstone (2012), and Ma (2013); these have been shown to attain the minimax rate under certain conditions, provided that the initial starting point is reasonably well-aligned with the true signal. We also mention the computationally-efficient combinatorial approaches proposed by Moghaddam, Weiss and Avidan (2006) and d'Aspremont, Bach and El Ghaoui (2008) that aim to find solutions to the optimization problem in (2) using greedy methods.

A common feature to all of the computationally efficient algorithms mentioned above is that they are iterative, in the sense that, starting from an initial guess $\hat{v}^{[0]} \in \mathbb{R}^p$, they refine their guess by producing a finite sequence of iterates $\hat{v}^{[1]}, \ldots, \hat{v}^{[T]} \in \mathbb{R}^p$, with the estimator defined to be the final iterate. A major drawback of such iterative methods is that a bad initialization may yield a disastrous final estimate. To illustrate this point, we ran a simple simulation in which the underlying distribution is $N_{100}(0, \Sigma)$, with

$$\Sigma = \begin{pmatrix} J_{10} & & \\ & 0.6J_{20} & \\ & & 1.1I_{70} \end{pmatrix},$$

where $J_q = \mathbf{1}_q\mathbf{1}_q^\top \in \mathbb{R}^{q\times q}$ denotes the matrix of ones. In this example, $v_1 = (\mathbf{1}_{10}^\top, \mathbf{0}_{90}^\top)^\top/\sqrt{10}$, so $k = 10$. Figure 1 shows, for several different SPCA algorithms and several different sample sizes, the average values of the loss function

$$L(u, v) := \sin \measuredangle(u, v) = \{1 - (u^\top v)^2\}^{1/2}, \tag{3}$$

over 100 repetitions of the experiment. Remarkably, each of the previously proposed algorithms we tested produces estimates that are almost orthogonal to the true principal component! The reason for this is that all of the default initialization procedures are unsuccessful in finding a good starting point; cf. Section 4.3 for further details.

In Section 2 of this paper, we propose a novel algorithm for SPCA that aggregates estimates over carefully-chosen random projections of the data into a lower-dimensional space. In contrast to the other algorithms mentioned above, it is non-iterative and does not depend on a choice of initialization, so it has no difficulty with the simulation example above; see the blue curve in Figure 1. Moreover, our algorithm, which we refer to as SPCAvRP and implement in a publicly available R package, is also attractive for both theoretical and computational reasons. Our theory, developed in Section 3, provides a detailed description of the statistical and computational trade-off involved in the SPCAvRP algorithm. It reveals a subtle interaction between conditions on an effective sample size parameter and the number of

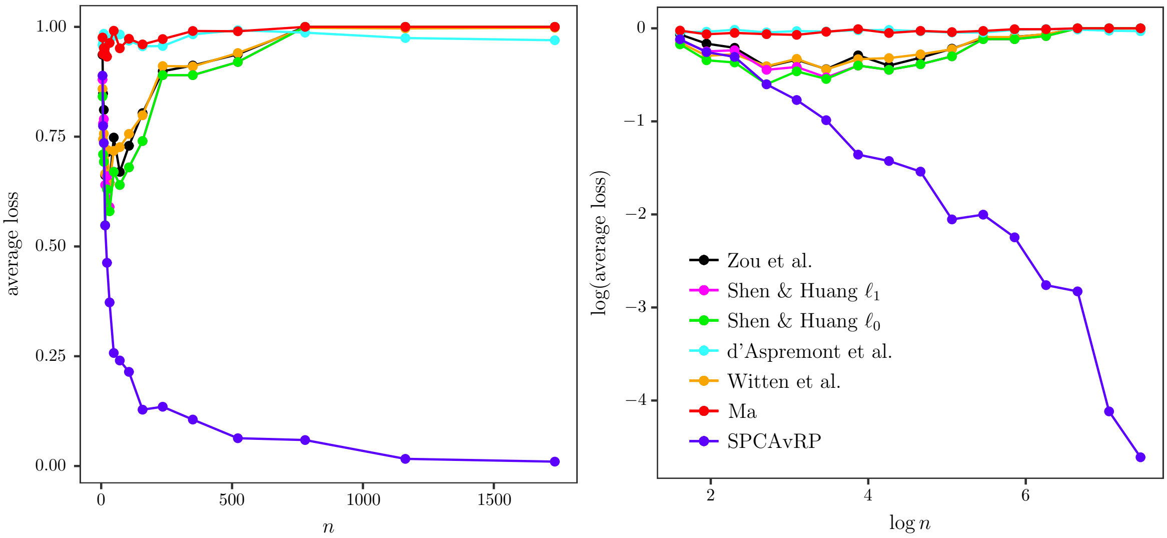


Figure 1: Average loss (3) (left) and its logarithm (right) for different sample sizes n. Blue: the SPCAvRP algorithm proposed in this paper; red: Ma (2013); orange: Witten, Tibshirani and Hastie (2009); cyan: d'Aspremont, Bach and El Ghaoui (2008); magenta and green: Shen and Huang (2008) with ℓ_1 and ℓ_0-thresholding; black: Zou, Hastie and Tibshirani (2006).

projections, under which our estimator attains the minimax optimal rate. When the effective sample size is large, the minimax rate can be attained with a number of projections that grows only slightly faster than linearly in p. This turns out not to contradict the computational lower bound of Wang, Berthet and Samworth (2016a), which applies to an intermediate effective sample size regime where the SPCAvRP algorithm would require an exponential number of projections to attain the optimal rate. The computational attractions of the proposed algorithm include the fact that it is embarrassingly parallelizable, and does not even require computation of $\hat{\Sigma} \in \mathbb{R}^{p\times p}$, since it suffices to extract principal submatrices of $\hat{\Sigma}$, which can be done by computing the sample covariance matrices of the projected data. This may result in a significant computational saving if p is very large. Several numerical aspects of the algorithm, including a finite-sample simulation comparison with alternative methods on both simulated and real data, are considered in Section 4. These reveal that our SPCAvRP algorithm has very competitive performance, and enjoys robustness properties that iterative algorithms do not share. The proofs of all of our results are given in Appendix A.

Algorithms based on random projections have recently been shown to be highly effective for several different problems in high-dimensional statistical inference. For instance, in the context of high-dimensional classification, Cannings and Samworth (2017) showed that their random projection ensemble classifier that aggregates over projections that yield small estimates of the test error can result in excellent performance. Marzetta, Tucci and Simon (2011) employ an ensemble of random projections to construct an estimator of the population covariance matrix and its inverse in the setting where $n < p$. Fowler (2009) introduced a so-called compressive-projection PCA that reconstructs the sample principal components from many low-dimensional projections of the data. Finally, to decrease the computational burden of classical PCA, Qi and Hughes (2012) and Pourkamali-Anaraki and Hughes (2014) propose estimating $v_1(\Sigma)$ by the leading eigenvector of $n^{-1}\sum_{i=1}^n P_i X_i X_i^\top P_i$, where $P_1, \ldots, P_n$ are random projections of a particular form.

Notation. We conclude this introduction with some notation used throughout the paper. For a vector $u \in \mathbb{R}^p$, we write $u^{(j)}$ for its jth component and let $\|u\| := \bigl\{\sum_{j=1}^p (u^{(j)})^2\bigr\}^{1/2}$ denote its Euclidean norm. For a real symmetric matrix $U \in \mathbb{R}^{p\times p}$, we let $\lambda_1(U) \geq \lambda_2(U) \geq \ldots \geq \lambda_p(U)$ denote its eigenvalues, arranged in decreasing order. In addition, we define the leading eigenvector of U by

$$v_1(U) := \operatorname*{sargmax}_{v\in\mathcal{S}^{p-1}} v^\top U v,$$

where sargmax denotes the smallest element of the argmax in the lexicographic ordering. In the special case where $U = \Sigma$, we drop the argument, and write the eigenvalues and eigenvectors as $\lambda_r = \lambda_r(\Sigma)$ and $v_r = v_r(\Sigma)$, respectively. We also define $U^{(j,j')}$ to be the (j,j')th entry of U, and write $\|U\|_{\mathrm{op}} := \sup_{\|x\|=1}\|Ux\|$ for the operator norm of matrix U.

For $r \in \mathbb{N}$, let $[r] := \{1, \ldots, r\}$. Let

$$S_1 = S_1(v_1) := \{j \in [p] : v_1^{(j)} \neq 0\}$$

denote the support of the vector $v_1 \in \mathbb{R}^p$. We write $v_1^{\min} := \min_{j\in S_1}|v_1^{(j)}|$ for the smallest non-zero component of v_1 in absolute value.

For any index subset $S \subseteq [p]$ we write P_S to denote the projection onto the span of $\{e_j : j \in S\}$, where $e_1, \ldots, e_p$ are the standard Euclidean basis vectors in $\mathbb{R}^p$, so that P_S is a $p \times p$ diagonal matrix whose jth diagonal entry is $\mathbb{1}_{\{j\in S\}}$. Finally, for $a, b \in \mathbb{R}$, we write $a \lesssim b$ to mean that there exists a universal constant $C > 0$ such that $a \leq Cb$.

2 SPCA via random projections

2.1 Single principal component estimation

In this section, we describe our algorithm for estimating a single principal component in detail; more general estimation of multiple principal components and principal subspaces is treated in Section 2.2 below. Let $x_1, \ldots, x_n$ be data points in $\mathbb{R}^p$ and let $\hat{\Sigma} := n^{-1}\sum_{i=1}^n x_i x_i^\top$. We think of $x_1, \ldots, x_n$ as independent realizations of a mean-zero random vector X, so a practitioner may choose to center each variable so that $\sum_{i=1}^n x_i^{(j)} = 0$ for each $j \in [p]$. For $d \in [p]$, let $\mathcal{P}_d := \{P_S : S \subseteq [p],\ |S| = d\}$ denote the set of d-dimensional, axis-aligned projections. For fixed $A, B \in \mathbb{N}$, consider projections $\{P_{a,b} : a \in [A], b \in [B]\}$ independently and uniformly distributed on $\mathcal{P}_d$. We think of these projections as consisting of A groups, each of cardinality B. For each $a \in [A]$, let

$$b^*(a) := \operatorname*{sargmax}_{b\in[B]} \lambda_1(P_{a,b}\hat{\Sigma}P_{a,b}).$$

The idea is that the non-zero entries of $P_{a,b^*(a)}\hat{\Sigma}P_{a,b^*(a)}$ form a principal submatrix of $\hat{\Sigma}$ that should have a large leading eigenvalue, so the non-zero entries of the corresponding leading eigenvector $\hat{v}_{a,b^*(a)}$ of $P_{a,b^*(a)}\hat{\Sigma}P_{a,b^*(a)}$ should have some overlap with those of v_1. Observe that, if $d = k$ and $\{P_{a,b} : b \in [B]\}$ contained all $\binom{p}{k}$ projections, then the leading eigenvector of $P_{a,b^*(a)}\hat{\Sigma}P_{a,b^*(a)}$ would yield the minimax optimal estimator in (2). Of course, it would typically be too computationally expensive to compute all such projections, so instead we only consider B randomly chosen ones.

The remaining challenge is to aggregate over the selected projections. To this end, for each coordinate $j \in [p]$, we compute the average $\hat{w}^{(j)}$ of the A absolute values of the jth components of the selected eigenvectors $\hat{v}_{a,b^*(a)}$. This means that we take account not just of the frequency with which each coordinate is chosen, but also their corresponding magnitudes in the selected eigenvector. Finally, we select the ℓ indices $\hat{S}_1$ corresponding to the largest values of $\hat{w}^{(1)}, \ldots, \hat{w}^{(p)}$ and output our estimate $\hat{v}_1$ as the leading eigenvector of $P_{\hat{S}_1}\hat{\Sigma}P_{\hat{S}_1}$. Pseudo-code for our SPCAvRP algorithm is given in Algorithm 1.

We remark that, by computing $\hat{w}$ in equation (5), the SPCAvRP algorithm ranks all of the coordinates according to their importance. As we shall see later in Section 4, this ranking turns out to be useful when choosing a suitable sparsity level ℓ for the final estimator in cases where the true sparsity level k is unknown in advance.

Algorithm 1: Pseudo-code for the SPCAvRP algorithm

```
Input: $x_1, \ldots, x_n \in \mathbb{R}^p$, $A, B \in \mathbb{N}$, $d, \ell \in [p]$.
Generate $\{P_{a,b} : a \in [A], b \in [B]\}$ independently and uniformly from $\mathcal{P}_d$.
Compute $\{P_{a,b}\hat{\Sigma}P_{a,b} : a \in [A], b \in [B]\}$, where $\hat{\Sigma} := n^{-1}\sum_{i=1}^n x_i x_i^\top$.
for $a = 1, \ldots, A$ do
    for $b = 1, \ldots, B$ do
        Compute $\hat{\lambda}_{a,b} := \lambda_1(P_{a,b}\hat{\Sigma}P_{a,b})$ and $\hat{v}_{a,b} \in v_1(P_{a,b}\hat{\Sigma}P_{a,b})$.
    end
    Compute
        $b^*(a) := \operatorname{sargmax}_{b \in [B]} \hat{\lambda}_{a,b}.$        (4)
end
Compute $\hat{w} = (\hat{w}^{(1)}, \ldots, \hat{w}^{(p)})^\top$, where
        $\hat{w}^{(j)} := \frac{1}{A}\sum_{a=1}^{A} |\hat{v}_{a,b^*(a)}^{(j)}|,$        (5)
  and let $\hat{S}_1 \subseteq [p]$ be the index set of the $\ell$ largest components of $\hat{w}$.
Output: $\hat{v}_1 := \operatorname{sargmax}_{v \in \mathcal{S}^{p-1}} v^\top P_{\hat{S}_1}\hat{\Sigma}P_{\hat{S}_1} v$.
```

Besides the intuitive selection of the most important coordinates, the use of axis-aligned projections in SPCAvRP algorithm facilitates faster computation as opposed to the use of general orthogonal projections. Indeed, the multiplication of $\hat{\Sigma} \in \mathbb{R}^{p\times p}$ by an axis-aligned projection $P \in \mathcal{P}_d$ from the left (or right) can be recast as the selection of d rows (or columns) of $\hat{\Sigma}$ corresponding to the indices of the non-zero diagonal entries of P. Thus, instead of the typical $\mathcal{O}(p^2 d)$ matrix multiplication complexity, only $\mathcal{O}(pd)$ operations are required. We also remark that, instead of storing P, it suffices to store its non-zero indices.

More generally, the computational complexity of Algorithm 1 can be analyzed as follows. Generating AB initial random projections takes $\mathcal{O}(ABd)$ operations. Next, we need to compute $P_{a,b}\hat{\Sigma}P_{a,b}$ for all a and b, which can be done in two different ways. One option is to compute $\hat{\Sigma}$, and then for each projection $P_{a,b}$ select the corresponding $d \times d$ principal submatrix of $\hat{\Sigma}$, which requires $\mathcal{O}(np^2 + ABd^2)$ operations. Alternatively, we can avoid computing $\hat{\Sigma}$ by computing the sample covariance matrix of the projected data $\{P_{a,b}x_1, \ldots, P_{a,b}x_n : a \in [A], b \in [B]\}$, which has $\mathcal{O}(ABnd^2)$ complexity. If $p^2 \gg ABd^2$, then the second option is preferable.

The rest of Algorithm 1 entails computing an eigendecomposition of each $d\times d$ matrix, and computing $b^*(a)$, $\hat{w}$, $\hat{S}_1$, and $\hat{v}_1$, which all together amounts to $\mathcal{O}(ABd^3 + p + \ell^3)$ operations. Thus, assuming that $n \geq d$, the overall computational complexity of the SPCAvRP algorithm is

$$\mathcal{O}(\min\{np^2 + ABd^3 + \ell^3, ABnd^2 + p + \ell^3\}).$$

We also note that, due to the use of random projections, the algorithm is highly parallelizable. In particular, both for-loops of Algorithm 1 can be parallelized, and the selection of good projections can easily be carried out using different (up to A) machines.

Finally, we note that the number of projections A and B, the dimension of those projections d, and the sparsity of the final estimator ℓ, need to be provided as inputs to Algorithm 1. The selection of these parameters is discussed in Section 4.2.

2.2 Multiple principal component estimation

The estimation of higher-order principal components is typically achieved via a deflation scheme. Having computed estimates $\hat{v}_1, \ldots, \hat{v}_{r-1}$ of the top $r-1$ principal components, the aim of such a procedure is to modify the observations $x_1, \ldots, x_n$ to remove correlation with these previously-estimated components (e.g. Mackey, 2009). Defining the random $p \times r$ matrix $\hat{V}_{r-1} := (\hat{v}_1, \ldots, \hat{v}_{r-1})$, one possibility is to set $\tilde{x}_i := H_{r-1}x_i$ for $i = 1, \ldots, n$, where $H_{r-1} := I_p - \hat{V}_{r-1}(\hat{V}_{r-1}^\top \hat{V}_{r-1})^{-1}\hat{V}_{r-1}^\top$ denotes the projection onto the orthogonal complement of the column space of $\hat{V}_{r-1}$. Note that, in contrast to classical PCA, in sparse PCA the estimated principal components from such a deflation scheme are typically not orthogonal. In Algorithm 2, we therefore propose a modified deflation scheme, which in combination with Algorithm 1 can be used to compute an arbitrary $s \in [p]$ principal components that are orthogonal (as well as sparse), as verified in Lemma 1 below.

Lemma 1. *For any $s \in [p]$, the outputs $\hat{v}_1, \ldots, \hat{v}_s$ of Algorithm 2 are mutually orthogonal.*

We remark that, in fact, our proposed deflation method can be used in conjunction with any SPCA algorithm.

Algorithm 2: Pseudo-code of the modified deflation scheme

```
Input: $x_1, \ldots, x_n \in \mathbb{R}^p$, $A, B \in \mathbb{N}$, $s, d, \ell_1, \ldots, \ell_s \in [p]$.
Let $\hat{v}_1$ be output of Algorithm 1 with inputs $x_1, \ldots, x_n$, $A$, $B$, $d$ and $\ell_1$.
for $r = 2, \ldots, s$ do
    Let $H_r := I_p - \hat{V}_{r-1}\hat{V}_{r-1}^\top$, where $\hat{V}_{r-1} := (\hat{v}_1, \ldots, \hat{v}_{r-1})$.
    Let $\tilde{v}_r$ be output of Algorithm 1 with inputs $H_r x_1, \ldots, H_r x_n$, $A$, $B$, $d$ and $\ell_r$.
    Let $\tilde{S}_r := \{j \in [p] : \tilde{v}_r^{(j)} \neq 0\}$ and $H_{\tilde{S}_r} := I_p - P_{\tilde{S}_r}\hat{V}_{r-1}(\hat{V}_{r-1}^\top P_{\tilde{S}_r}\hat{V}_{r-1})^{-1}\hat{V}_{r-1}^\top P_{\tilde{S}_r}$.
    Compute
        $\hat{v}_r := v_1\big(H_{\tilde{S}_r} P_{\tilde{S}_r} \hat{\Sigma} P_{\tilde{S}_r} H_{\tilde{S}_r}\big).$
end
Output: $\hat{v}_1, \ldots, \hat{v}_s$.
```

Although Algorithm 2 can conveniently be used to compute sparse principal components up to order s, it requires Algorithm 1 to be executed s times. Instead, we can modify Algorithm 1 to estimate directly the leading eigenspace of dimension s at a considerably reduced computational cost. To this end, we propose a generalization of the SPCAvRP algorithm for eigenspace estimation in Algorithm 3. In this generalization, $s \times A$ projections are selected from total of $A \times B$ random projections, by computing

$$b_r^*(a) := \operatorname*{sargmax}_{b \in [B]} \lambda_r(P_{a,b}\hat{\Sigma}P_{a,b})$$

for each $r \in [s]$ and $a \in [A]$. Moreover, the aggregation step is modified to account for the sparsity and orthogonality of the components. Observe that for $s = 1$, Algorithm 3 reduces to Algorithm 1. Furthermore, for any s, up to the step where $b_r^*(a)$ is computed, Algorithm 3 has the same complexity as Algorithm 1, with the total complexity of Algorithm 3 amounting to $\mathcal{O}(\min\{np^2 + ABd^3 + s^4\ell^3, ABnd^2 + sp + s^4\ell^3\})$ provided that $n \geq d$, $\min(d^3, Bd^2) \geq s$ and $\ell_r = \mathcal{O}(\ell)$.

3 Theoretical guarantees

In this section, we focus on Algorithm 1 and assume that $X_1, \ldots, X_n$ are independently sampled from a distribution Q satisfying a Restricted Covariance Concentration (RCC) condition introduced in Wang, Berthet and Samworth (2016a). Recall that, for $K > 0$, we say that a mean zero distribution Q on $\mathbb{R}^p$ satisfies an RCC condition with parameter K, and write $Q \in \mathrm{RCC}_p(K)$, if for all $\delta > 0$, $n \in \mathbb{N}$ and $r \in \{1, \ldots, p\}$, we have

$$\mathbb{P}\Bigg\{\sup_{u \in \mathcal{B}_0^{p-1}(r)} \big|u^\top(\hat{\Sigma} - \Sigma)u\big| \geq K \max\Bigg(\sqrt{\frac{r\log(p/\delta)}{n}}, \frac{r\log(p/\delta)}{n}\Bigg)\Bigg\} \leq \delta. \qquad (6)$$

Algorithm 3: Pseudo-code of the SPCAvRP algorithm for eigenspace estimation

```
Input: $x_1, \dots, x_n \in \mathbb{R}^p$, $A, B \in \mathbb{N}$, $s, d, \ell_1, \dots, \ell_s \in [p]$.
Generate $\{P_{a,b} : a \in [A], b \in [B]\}$ independently and uniformly from $\mathcal{P}_d$.
Compute $\{P_{a,b}\hat{\Sigma}P_{a,b} : a \in [A], b \in [B]\}$, where $\hat{\Sigma} := n^{-1}\sum_{i=1}^n x_i x_i^\top$.
for $a = 1, \dots, A$ do
    for $b = 1, \dots, B$ do
        for $r = 1, \dots, s$ do
            Compute $\hat{\lambda}_{a,b;r} := \lambda_r(P_{a,b}\hat{\Sigma}P_{a,b})$ and the corresponding eigenvector $\hat{v}_{a,b;r}$.
        end
    end
    for $r = 1, \dots, s$ do
        $b_r^*(a) := \mathrm{sargmax}_{b \in [B]} \hat{\lambda}_{a,b;r}$.
    end
end
for $r = 1, \dots, s$ do
    Let $\hat{w}_r = (\hat{w}_r^{(1)}, \dots, \hat{w}_r^{(p)})^\top$ be such that $\hat{w}_r^{(j)} := \frac{1}{A}\sum_{a=1}^A \bigl|\hat{v}^{(j)}_{a,b_r^*(a);r}\bigr|$ and let $\hat{S}_r^{\mathrm{init}}$ be
      the index set of the $\sum_{q=1}^r \ell_q$ largest components in $\hat{w}_r$.
    Let $H_r := I_p - \hat{V}_{r-1}\hat{V}_{r-1}^\top$, where $\hat{V}_{r-1} := (\hat{v}_1, \dots, \hat{v}_{r-1})$, and let $\hat{S}_r$ be the index set
      of the $\ell_r$ largest absolute values of the components of $v_1\bigl(P_{\hat{S}_r^{\mathrm{init}}} H_r \hat{\Sigma} H_r P_{\hat{S}_r^{\mathrm{init}}}\bigr)$.
    Let $H_{\hat{S}_r} := I_p - P_{\hat{S}_r}\hat{V}_{r-1}(\hat{V}_{r-1}^\top P_{\hat{S}_r}\hat{V}_{r-1})^{-1}\hat{V}_{r-1}^\top P_{\hat{S}_r}$ and compute
        $$\hat{v}_r := v_1\bigl(H_{\hat{S}_r} P_{\hat{S}_r} \hat{\Sigma} P_{\hat{S}_r} H_{\hat{S}_r}\bigr).$$
end
Output: $\hat{v}_1, \dots, \hat{v}_s$.
```

In particular, if $Q = N_p(0, \Sigma)$, then $Q \in \mathrm{RCC}_p\bigl(8\lambda_1(1 + 9/\log p)\bigr)$; and if Q is sub-Gaussian with parameter σ^2, in the sense that $\int_{\mathbb{R}^p} e^{u^\top x}\, dQ(x) \leq e^{\sigma^2\|u\|^2/2}$ for all $u \in \mathbb{R}^p$, then $Q \in \mathrm{RCC}_p\bigl(16\sigma^2(1 + 9/\log p)\bigr)$ (Wang, Berthet and Samworth, 2016a, Proposition 1). In Section 3.1, we first derive theoretical guarantees in the special case where the covariance matrix Σ has a single-spiked structure and its leading eigenvector is homogeneous in all signal coordinates. The result in this special case already provides useful insights on how different parameters affect the performance of estimator proposed in Algorithm 1. We then extend our theory to more general distributions in Section 3.2.

3.1 Single-spiked model with homogeneous signal

Any permutation π of $[p]$ acts naturally on $\mathbb{R}^p$ by $\pi(x^{(1)}, \ldots, x^{(p)}) = (x^{(\pi(1))}, \ldots, x^{(\pi(p))})$. This action maps any probability measure Q on $\mathbb{R}^p$ to another probability measure $\pi_* Q$ on $\mathbb{R}^p$, where for any Borel set $A \subseteq \mathbb{R}^p$, we define $\pi_* Q(A) = Q(\pi(A))$. In this section, we consider a subclass of distributions

$$\mathcal{Q}_0 \subseteq \mathrm{RCC}_p(K)$$

such that any $Q \in \mathcal{Q}_0$ has covariance matrix $\Sigma = I_p + \theta_1 v_1 v_1^\top$, for some $\theta_1 > 0$ and $v_1 := k^{-1/2}(\mathbf{1}_k^\top, \mathbf{0}_{p-k}^\top)^\top \in \mathcal{B}_0^{p-1}(k)$ and such that $Q = \pi_* Q$ for any π that stabilizes $\{1, \ldots, k\}$, that is $\{1, \ldots, k\} = \{\pi(1), \ldots, \pi(k)\}$. In particular, $\mathcal{Q}_0$ includes distributions of the form $N_p(0, I_p + \theta_1 v_1 v_1^\top)$ when $K \geq 8(1 + \theta_1)(1 + 9/\log p)$ and $v_1 = k^{-1/2}(\mathbf{1}_k^\top, \mathbf{0}_{p-k}^\top)^\top$.

In what follows, we use $F_{\mathrm{HG}}(\cdot; d, k, p)$ to denote the distribution function of the hypergeometric distribution $\mathrm{HyperGeom}(d, k, p)$. Recall that this distribution models the number of white balls obtained when drawing d balls uniformly and without replacement from an urn containing p balls, k of which are white.

Theorem 2. *Let $X_1, \ldots, X_n \overset{\mathrm{iid}}{\sim} Q \in \mathcal{Q}_0$. Let $\hat{v}_1$ be the output of Algorithm 1 with input $X_1, \ldots, X_n$, A, B, d and ℓ. Assume that $p \geq \max(4, 2k)$, $n \geq 4\max(d, \ell)\log p$, and that there exists $t \in \{1, \ldots, k\}$ such that*

$$\bigl\{1 - F_{\mathrm{HG}}(t-1; d, k, p)\bigr\} B \geq 3\log p \tag{7}$$

and

$$40K\sqrt{\frac{k^2 d \log p}{t^2 n \theta_1^2}} \leq \min\{1, (p-k)d^{-1/2}k^{-1}\}. \tag{8}$$

Then with probability at least $1 - p^{-3} - pe^{-A/(32k^2)}$ we have

$$L(\hat{v}_1, v_1) \leq 4K\sqrt{\frac{\ell \log p}{n\theta_1^2}} \max\Bigl(1, \frac{k}{\ell}\Bigr) + \sqrt{\max\Bigl(1 - \frac{\ell}{k}, 0\Bigr)}.$$

We note that for $\ell \geq k$, the loss is bounded by a constant multiple of $\sqrt{\ell \log p/(n\theta_1^2)}$, which, as mentioned in the introduction, is minimax rate optimal when ℓ/k is bounded by a constant. On the other hand, when $\ell < k$, we may incur an additional loss of order $\sqrt{1 - \ell/k}$. This conclusion exhibits the trade-off in the choice of ℓ, which is also conveyed by the numerical examples in Section 4.

As t increases, conditions (7) and (8) are respectively strengthened and weakened. The flexibility of t allows us to trade off these two conditions in the theorem. For example, when $t = 1$, we have

$$F_{\mathrm{HG}}(0; d, k, p) = \frac{\binom{p-k}{d}}{\binom{p}{d}} \leq 1 - k/p,$$

so it suffices to choose $B \geq 3k^{-1}p \log p$ for (7) to hold. In this case, we can choose parameters A and B, depending polynomially on p and k, so that Algorithm 1 is a polynomial time algorithm that can achieve the minimax rate for $\theta_1 \leq 1$ and appropriately chosen ℓ. However, this does not contradict the computational lower bound established in Wang, Berthet and Samworth (2016a, Theorem 6) because for $t = 1$, condition (8) implies a sample size requirement of order $n \gtrsim K^2 k^2 d \theta_1^{-2} \log p$, which belongs to the high effective sample size regime discussed in Wang, Berthet and Samworth (2016a, Section 4.4). On the other hand, for $t \asymp k$, (8) is satisfied for a much smaller sample size $n \gtrsim K^2 d \theta_1^{-2} \log p$, which includes both the intermediate and high effective sample size regimes of Wang, Berthet and Samworth (2016a) (these are the only regimes where consistent estimation is possible using any algorithm). However, by Hoeffding (1963, Theorems 2 and 4), if $t \geq dk/p$, then

$$1 - F_{\mathrm{HG}}(t - 1; d, k, p) \leq \exp\{-2d(t - dk/p)^2\},$$

which together with (7) entails choosing B exponentially large in the problem parameters. Hence Algorithm 1 will not be polynomial time in this case. Therefore, in this single-spiked homogeneous signal setting, Theorem 2 continuously interpolates between the high and intermediate effective sample size regimes, and elucidates the phase transition for our random projection ensemble estimator in a fairly precise way.

3.2 General distributions

We consider more general distributions in this section. To begin with, we provide a proposition which controls the risk of estimator $\hat{v}_1$ defined in Algorithm 1 by the sum of a bias term, based on its support recovery quality, and a variance term, which measures the risk incurred in estimating the leading eigenvector after knowing its support.

Proposition 3. *Let $X_1, \ldots, X_n \overset{\mathrm{iid}}{\sim} Q \in \mathrm{RCC}_p(K)$ with covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$ satisfying $\lambda_1 > \lambda_2$. Suppose that $v_1 \in \mathcal{B}_0^{p-1}(k)$ has support S_1. Let $\hat{\Sigma} := n^{-1} \sum_{i=1}^n X_i X_i^\top$ and let $\hat{v}_1 = v_1\bigl(P_{\hat{S}_1} \hat{\Sigma} P_{\hat{S}_1}\bigr)$, where $\hat{S}_1 = \hat{S}_1(X_1, \ldots, X_n)$ is a random subset of $[p]$ of cardinality ℓ. If $p \geq 3$ and $n \geq \ell \log p$, then*

$$\mathbb{E} L(\hat{v}_1, v_1) \leq \frac{4K}{\lambda_1 - \lambda_2} \sqrt{\frac{\ell \log p}{n}} + \mathbb{P}(S_1 \not\subseteq \hat{S}_1).$$

Note that this result holds for any estimator of form $v_1\big(P_{\hat{S}_1}\hat{\Sigma}P_{\hat{S}_1}\big)$ where $\hat{S}_1$ is an index subset of cardinality ℓ that depends on the data. In what follows, we bound $\mathbb{P}\big(S_1 \not\subseteq \hat{S}_1\big)$ when $\hat{S}_1$ is defined as in Algorithm 1, thereby explicitly bounding the risk of estimator $\hat{v}_1$ computed therein. To achieve this, we show that with high probability, our selection criterion (4) ensures that we aggregate over a certain set of 'good' projections, defined for $\tau \in (0,1]$ by

$$\mathcal{G} = \mathcal{G}_\tau := \big\{P \in \mathcal{P}_d : \|Pv_1\| \geq \tau\big\}.$$

Such projections capture at least a given proportion of the signal in the leading eigenvector v_1. Writing $P_1 := P_{1,b^*(1)}$ for the selected projection from the first group of B projections in Algorithm 1, we also define the event

$$\Omega = \Omega_{\tau,B} := \{P_1 \in \mathcal{G}\}.$$

Since we aim to bound $\mathbb{P}\big(S_1 \not\subseteq \hat{S}_1\big)$, and since signal coordinates may differ in magnitude, we need to consider the probability that each signal coordinate $j \in S_1$ is captured by a selected good projection. To this end, we define

$$\rho := \max_{j \in S_1} \mathbb{P}\big(\big\{P_1^{(j,j)} = 1\big\} \cap \Omega\big) - \min_{j \in S_1} \mathbb{P}\big(\big\{P_1^{(j,j)} = 1\big\} \cap \Omega\big). \tag{9}$$

Observe that, under the setting of Section 3.1, we have $\rho = 0$. Moreover, whenever $v_1 \in \mathcal{B}_0^{(p-1)}(k)$ and $\mathbb{E}(\|X_1\|^2) < \infty$, we have $\lim_{n,B\to\infty} \rho = 0$ provided $d \geq k$. The theorem below provides conditions under which we can control $\mathbb{P}\big(S_1 \not\subseteq \hat{S}_1\big)$, and therefore bound the risk of our SPCAvRP estimator. In what follows, we order $(v_1^{(1)})^2, \ldots, (v_1^{(p)})^2$ as $v_{1,(1)}^2 \geq \cdots \geq v_{1,(p)}^2$.

Theorem 4. *Let $X_1, \ldots, X_n \overset{\mathrm{iid}}{\sim} Q \in \mathrm{RCC}_p(K)$ with covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$ satisfying $\lambda_1 > \lambda_2$ and $v_1 \in \mathcal{B}_0^{p-1}(k)$. Let $\hat{v}_1$ be the output of Algorithm 1 with input $X_1, \ldots, X_n$, A, B, d and ℓ, satisfying $\ell \geq k$ and $n \geq 4\max(d,\ell)\log p$ and $p \geq 3$. Suppose there exists $\tau \in (\sqrt{\lambda_2/\lambda_1}, 1]$ such that*

$$\varepsilon := \left\{\left(1 - \frac{2}{p^3}\right)\tau^2 - \rho\right\}v_1^{\min} - \frac{8\sqrt{2}K}{\lambda_1\tau^2 - \lambda_2}\sqrt{\frac{d\log p}{n}} - \frac{4\sqrt{2}\lambda_2}{\lambda_1\tau^2} - 2p^{-3} > 0, \tag{10}$$

and that B is large enough that there exists $k' \in [k]$ for which

$$\left\{1 - F_{\mathrm{HG}}\left(\frac{1}{v_{1,(k')}^2}\left(\tau^2 + \frac{4K}{\lambda_1}\sqrt{\frac{d\log p}{n}} + \frac{\lambda_2}{\lambda_1}\right); d, k', p\right)\right\}B \geq 3\log p. \tag{11}$$

Then

$$\mathbb{E}L(\hat{v}_1, v_1) \leq \frac{4K}{\lambda_1 - \lambda_2}\sqrt{\frac{\ell\log p}{n}} + pe^{-A\varepsilon^2/8}.$$

Remark: In the case where Σ is a spiked covariance matrix of the form

$$\Sigma = I_p + \sum_{r=1}^{m} \theta_r v_r v_r^\top, \tag{12}$$

for some $\theta_1 > \theta_2 \geq \cdots \geq \theta_m > 0$ and orthonormal vectors $v_1 \in \mathcal{B}_0^{p-1}(k)$, $v_2, \ldots, v_m \in \mathcal{S}^{p-1}$, the conditions of Theorem 4 can be weakened. In fact, noting the remarks following Lemma A.2 and Lemma A.3, λ_1 and λ_2 in the theorem may be replaced with θ_1 and θ_2 respectively (the naive direct application of Theorem 4 would have set $\lambda_r = 1 + \theta_r$ for $r = 1, 2$).

We further remark that conditions (10) and (11) again exhibit a statistical and computational trade-off as discussed after Theorem 2. For τ close 1, (10) is satisfied with a mild sample size requirement but (11) would require a choice of B exponentially large in the problem parameters. On the other hand, if λ_2/λ_1 is sufficiently small and τ is close to $\sqrt{\lambda_2/\lambda_1}$, then (11) can be satisfied with a B polynomial in the problem parameters, at the price of a much larger sample size requirement implied by (10).

4 Numerical experiments

In this section we demonstrate the performance of our proposed method on several numerical examples and discuss the choice of the different input parameters. We also compare our method with several existing sparse principal component estimation algorithms. All examples are computed using the R package 'SPCAvRP' (Gataric, Wang and Samworth, 2017).

4.1 Dependence of risk on problem parameters

Our first goal is to illustrate that our SPCAvRP algorithm achieves the estimation risk bounds as derived in Section 3. To this end, we apply Algorithm 1 to observations independently and identically sampled from a $N_p(0, \Sigma)$ distribution with a spiked covariance matrix Σ defined as in (12). It is convenient to define the effective sample size

$$n_{\mathrm{eff}} := \frac{n}{k \log p},$$

and in Figure 2, we plot the loss $L(\hat{v}_1, v_1)$, averaged over 100 repetitions for a range of values of n_{eff}. In addition to the empirical loss, we also plot $n_{\mathrm{eff}}^{-1/2}$ and an empirical estimate of $\mathbb{P}(S_1 \not\subseteq \hat{S}_1)$, which, up to universal scaling constants, are the two terms in the risk bound derived in Proposition 3. We observe that the curves of empirical losses for different values of p align well with each other, showing that n_{eff} is indeed an effective sample size that characterizes the difficulty of the estimation problem. We also observe that the empirical estimate for $\mathbb{P}(S_1 \not\subseteq \hat{S}_1)$ exhibits a rapid phase transition in its behavior as n_{eff} increases. Thus, for moderately large n_{eff}, the loss $L(\hat{v}_1, v_1)$ is essentially controlled by $n_{\mathrm{eff}}^{-1/2}$, which is reflected by the linear decay of the loss curve with slope $-1/2$ under the log-log scaling in Figure 2. In fact, since the left panel of Figure 2 corresponds to the single-spiked homogeneous signal setting, in this special case we can apply Theorem 2 to bound the risk by

$64(1 + 9\log^{-1} p)n_{\mathrm{eff}}^{-1/2}$ in the high effective sample size regime discussed after Theorem 2. However, we note that the loss curves behave very similarly in both panels, indicating that the algorithmic performance is robust to the presence of multiple spikes.

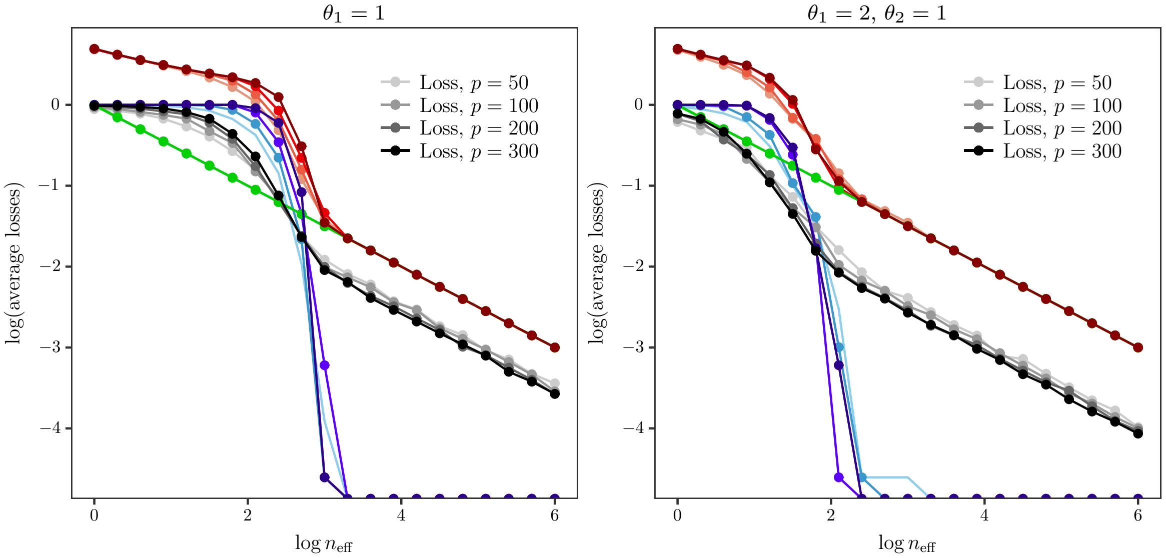


Figure 2: Estimation rate of the SPCAvRP algorithm for the Gaussian spiked model. Left panel: $\Sigma = I_p + \theta_1 v_1 v_1^\top$ for $v_1 = k^{-1/2}(\mathbf{1}_k^\top, \mathbf{0}_{p-k}^\top)^\top$; right panel: $\Sigma = I_p + \theta_1 v_1 v_1^\top + \theta_2 v_2 v_2^\top$ for $v_1 = k^{-1/2}(\mathbf{1}_k^\top, \mathbf{0}_{p-k}^\top)^\top$ and $v_2 = k^{-1/2}(\mathbf{0}_k^\top, \mathbf{1}_k^\top, \mathbf{0}_{p-2k}^\top)^\top$. Black: $L(\hat{v}_1, v_1)$ averaged over 100 experiments; blue: $\mathbb{P}(S_1 \nsubseteq \hat{S}_1)$ estimated over 100 experiments; green: $n_{\mathrm{eff}}^{-1/2}$; red: $n_{\mathrm{eff}}^{-1/2} + \mathbb{P}(S_1 \nsubseteq \hat{S}_1)$. Light to dark colors are for the choice of (p, A, B) corresponding to $(50, 200, 100)$, $(100, 300, 150)$, $(200, 400, 200)$ and $(300, 500, 250)$ respectively. In both panels, $d = \ell = k = \lfloor \sqrt{p} \rfloor$.

4.2 Choice of input parameters

4.2.1 Choice of A and B

We first consider the choice of parameters A and B, the number of groups of projections and the cardinality of each group respectively. Our SPCAvRP algorithm first selects the best projection in each group and then aggregates over all selected projections to obtain the final estimator. In Figure 3, we demonstrate that the selection step within each group of projections is crucial to the success of the algorithm. Specifically, we see that using the same total number of random projections, our two-stage procedure has superior performance over the naive aggregation over all projections, which corresponds to setting $B = 1$ in Algorithm 1. Interestingly, Figure 3 shows that simply increasing the number of projections, without performing a selection step, does not noticeably improve the performance of the basic aggregation. We note that even for the relatively small choices $A = 50$ and $B = 25$, the SPCAvRP algorithm does significantly better than the naive aggregation over 180000 projections.

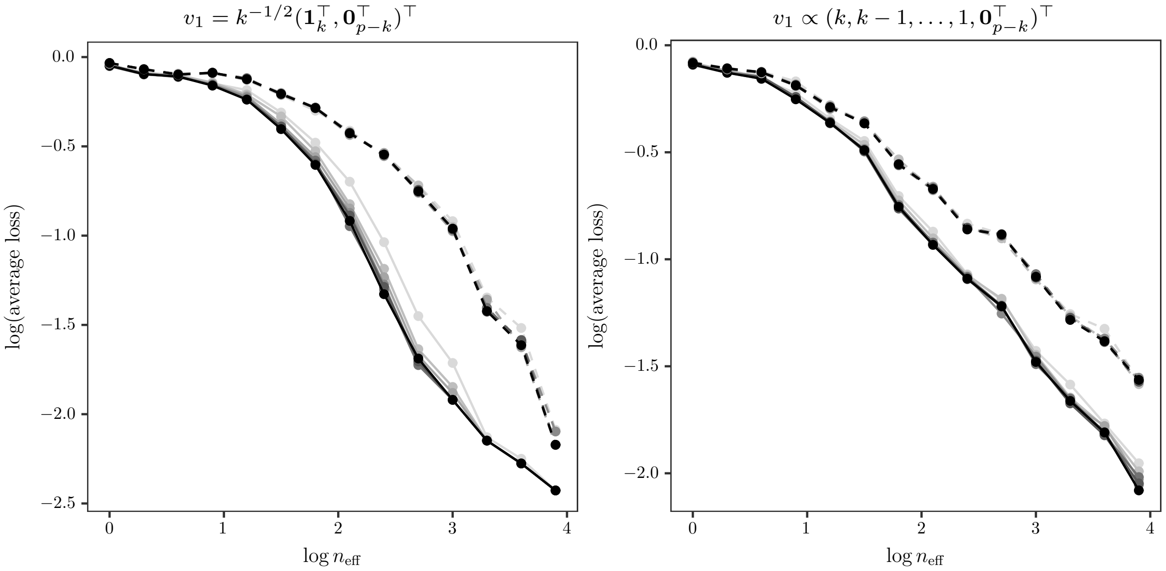


Figure 3: Solid lines, light to dark grey: (A, B) is $(50, 25)$, $(100, 50)$, $(200, 100)$, $(300, 150)$, $(400, 200)$, $(500, 250)$, $(600, 300)$. Dashed lines, light to dark grey: $B = 1$ and A is 50×25, 100×50, 200×100, 300×150, 400×200, 500×250, 600×300. In both panels, the distribution is $N_p(0, I_p + v_1 v_1^\top)$, $p = 50$, $d = l = k = \lfloor \sqrt{p} \rfloor = 7$.

Figure 4 demonstrates the effect of increasing either A or B while keeping the other fixed. We can see that increasing A and B noticeably improves the estimation quality in the medium effective sample size regime, and the benefit is more prominent when A and B are relatively small. Note that in Theorem 4, the risk bound improves as A increases. Thus, in practice, we would like to choose A as big as possible subject to our computational budget. The choice of B, however, is a little more delicate. In some settings, such as the single-spiked, homogeneous model in Figure 4 where the parameter ρ in (9) is zero, the performance appears to improve steadily as B increases. On the other hand, we can also construct examples where ρ appears to depend in a non-monotonic way on B, and performance does not necessarily improve as B increases; see Figure 5. In general, from our numerical experiments, we find that A and B should increase with p. We suggest using $A = 300$ and $B = 100$ when $p \approx 100$, while $A = 600$ and $B = 200$ when $p \approx 1000$.

4.2.2 Choice of d and ℓ

So far, in all our simulations, we have assumed that the true sparsity level k is known and we took $d = \ell = k$, where d is the dimension of the random projections and ℓ is the sparsity of the computed estimator. However, in practice k may not be known in advance. In Figure 6, we demonstrate how the over- and under-estimation of k affects the loss of our estimator. In particular, we choose $d = \ell = k \pm j$ for a range of values of j in both the homogeneous and inhomogeneous single-spiked Gaussian models. We incur a greater loss in under-estimating rather than over-estimating k, especially in the case when the signal coordinates are of

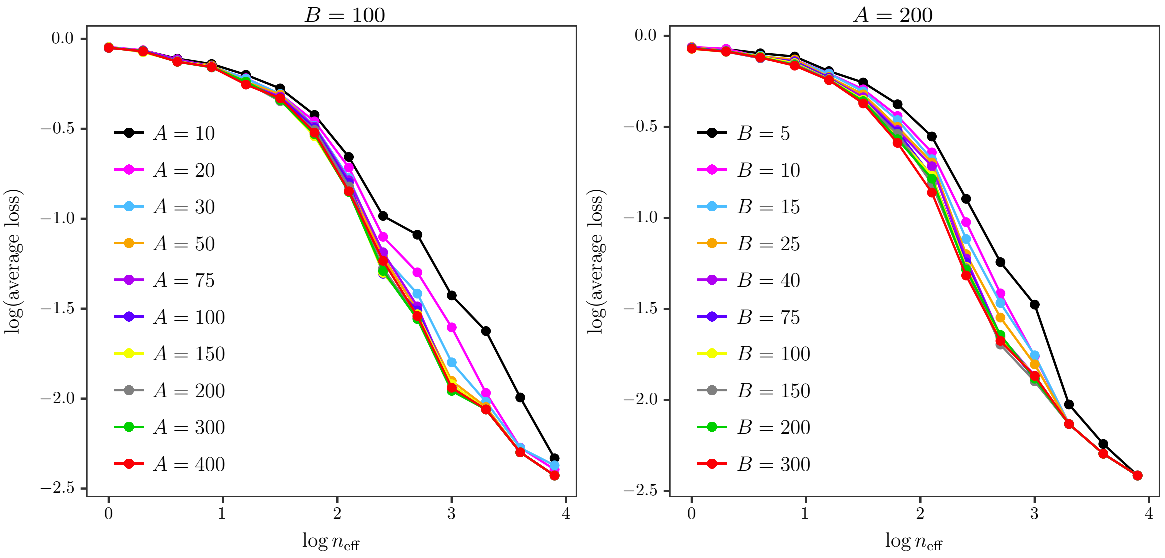


Figure 4: Average losses on the logarithmic scale as a function of the logarithm of the effective sample size n_{eff}. In the left panel, $B = 100$ and A is varied; on the right, $A = 200$ and B is varied. In both panels, the distribution is $N_p(0, I_p + v_1 v_1^\top)$ with $v_1 = k^{-1/2}(\mathbf{1}_k^\top, \mathbf{0}_{p-k}^\top)^\top$, $p = 50$, $d = l = k = \lfloor\sqrt{p}\rfloor = 7$.

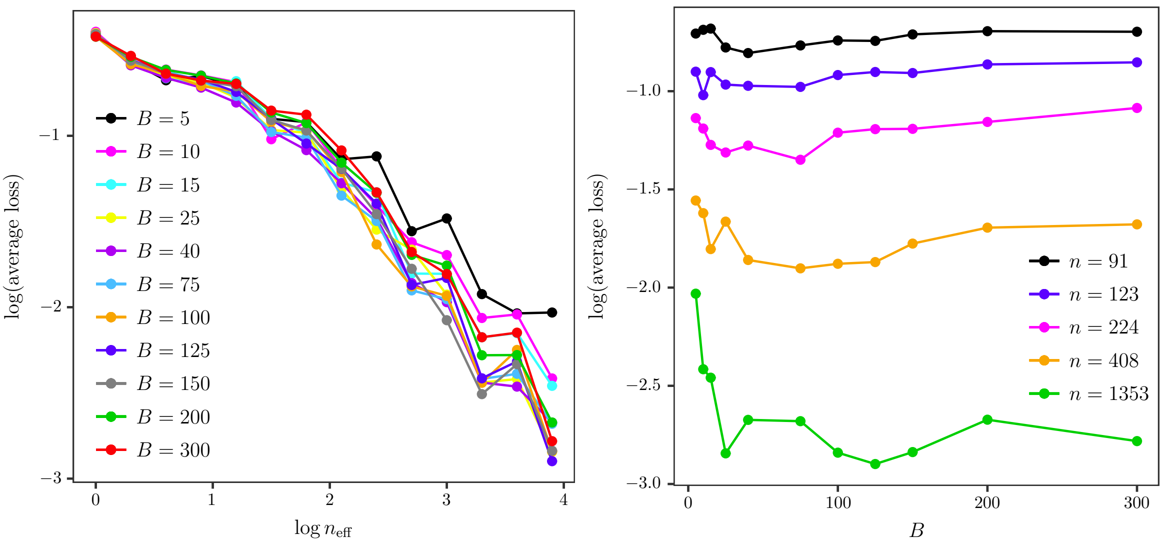


Figure 5: Trade-off in the choice of B. Left panel: the logarithm of average losses as a function of $\log n_{\text{eff}}$, where B is varied. Right panel: the logarithm of average losses as a function of B, where n is varied. In both panels, $A = 200$ and the distribution is $N_p(0, I_p + 5v_1 v_1^\top + 4v_2 v_2^\top)$ with $v_1 = k^{-1/2}(\mathbf{1}_k^\top, \mathbf{0}_{p-k}^\top)^\top$, $v_2 = k^{-1/2}(\mathbf{1}_3^\top, -1, 1, -1, 1, -1, 1, 1, \mathbf{0}_{p-k-3}^\top)^\top$, $p = 50$, $d = l = k = \lfloor\sqrt{p}\rfloor = 7$.

comparable magnitude, as shown in the left panel of Figure 6. In fact, it is interesting to note that, particularly in the medium effective sample size regime, over-estimating k may actually yield improved performance compared with using the true value of k; in this regime the term $\mathbb{P}(S_1 \not\subseteq \hat{S}_1)$ in the bound in Proposition 3 is not vanishingly small, and is reduced by choosing $d = \ell$ to be larger.

In addition, in Figure 7 we investigate the robustness of SPCAvRP to the choice of projection dimension d. We see that for a wide range of d values, the loss curves are close to each other. In fact, for homogeneous signal, the loss curves for different choices of d merge in the high effective sample size regime, whereas in the intermediate effective sample size regime, we may again see improved performance when d exceeds k. In the inhomogeneous case, the loss curves exhibit little dependence on d. In view of the above discussion regarding Figures 6 and 7, we conclude that our algorithm is more robust to the choice of d than to the choice of ℓ. In fact, large values of ℓ increase the chance that signal coordinates are discovered but also increase the probability of including noise coordinates. This trade-off in the choice of ℓ is also reflected in the bound derived in Theorem 2.

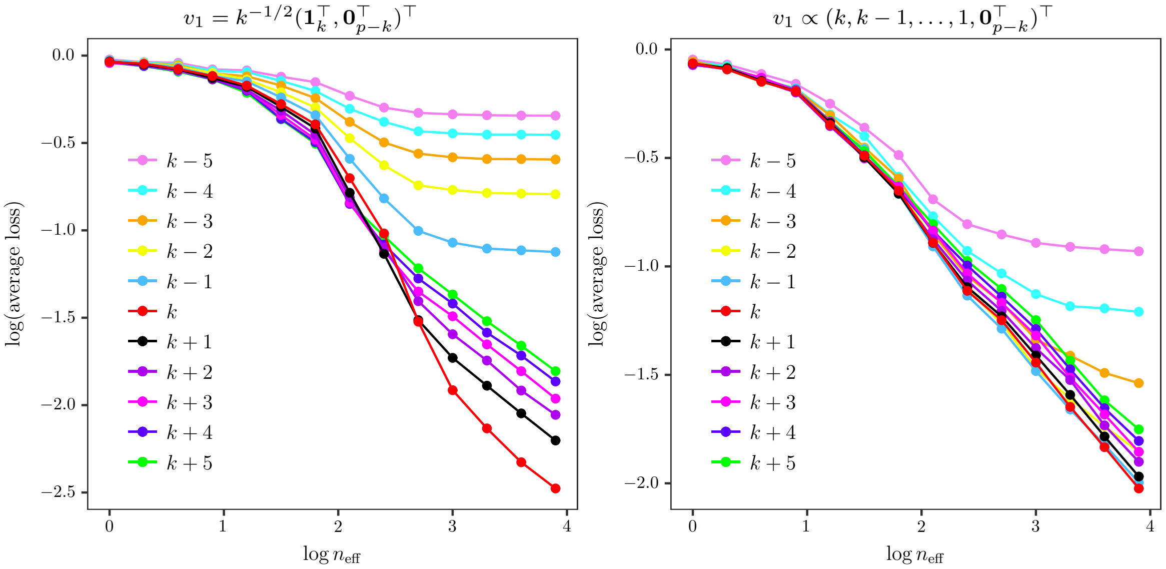
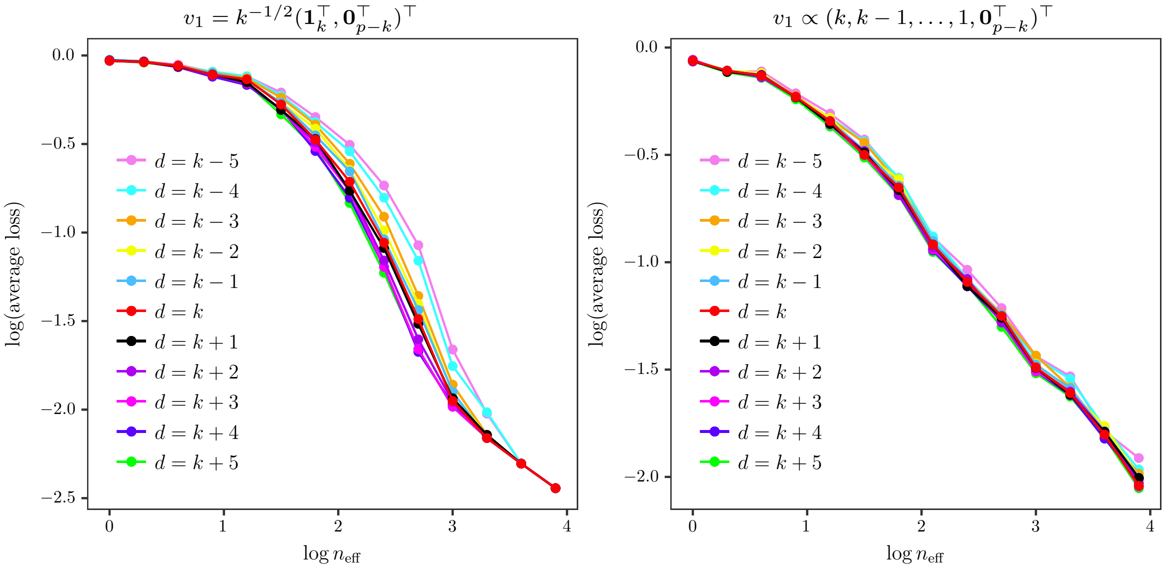


Figure 6: Over and under-shooting of k ($\ell = d = k \pm j$) for Gaussian single-spiked model with $\theta_1 = 1$, $p = 100$, $k = 10$, $A = 150$, $B = 50$.

In practice, when k is not given in advance, we would advocate a relatively small choice of d, given that this parameter has a high impact on the overall computational complexity and relatively small impact on the statistical performance of the algorithm. The choice of ℓ may be made by inspecting the total variance, a popular approach in the SPCA literature (e.g. Shen and Huang, 2008). More precisely, for each ℓ on a grid of plausible values, we can compute an estimate $\hat{v}_{1,\ell} \in \mathcal{B}_0(\ell)$ using the SPCAvRP algorithm and its explained variance $V_\ell := \hat{v}_{1,\ell}^\top \hat{\Sigma} \hat{v}_{1,\ell}$, and then plot V_ℓ against ℓ. As can be seen from Figure 8, the explained variance increases with ℓ, but plateaus off for $\ell \geq k$. An attractive feature of our method is that there is no need to re-run the entire algorithm for each value of ℓ. Recall that $\hat{w}$ in (5) of Algorithm 1 ranks the coordinates by their importance. Therefore, we only need to compute $\hat{w}$ once and then calculate the explained variance by selecting the top ℓ coordinates in $\hat{w}$ for each value of ℓ.

In cases where higher-order principal components need to be computed, a practical choice for $\ell_1, \dots, \ell_s$ in Algorithms 2 and 3 is $\ell_1 = \cdots = \ell_s = \ell$. However, it is also possible to choose each individual ℓ_r differently by inspecting the total variance at each iteration r of Algorithm 2.


$v_1 = k^{-1/2}(\mathbf{1}_k^\top, \mathbf{0}_{p-k}^\top)^\top$

$v_1 \propto (k, k-1, \dots, 1, \mathbf{0}_{p-k}^\top)^\top$

Figure 7: Choosing different d for Gaussian single-spiked model with $\theta_1 = 1$, $p = 100$, $k = 10$, $A = 150$, $B = 50$, and $\ell = k$.

4.3 Comparison with existing methods

In this subsection, we compare our method with several existing approaches for SPCA. We first present two simulated examples where only the first principal component is computed, and then also the examples of higher-order principal component estimation and an illustration on some genetic data.

4.3.1 First principal component

In addition to the example presented in Figure 1 of the introduction, we consider two further examples with multivariate Gaussian data generated via $Q = N_p(0, \Sigma)$ for $p = 100$ and

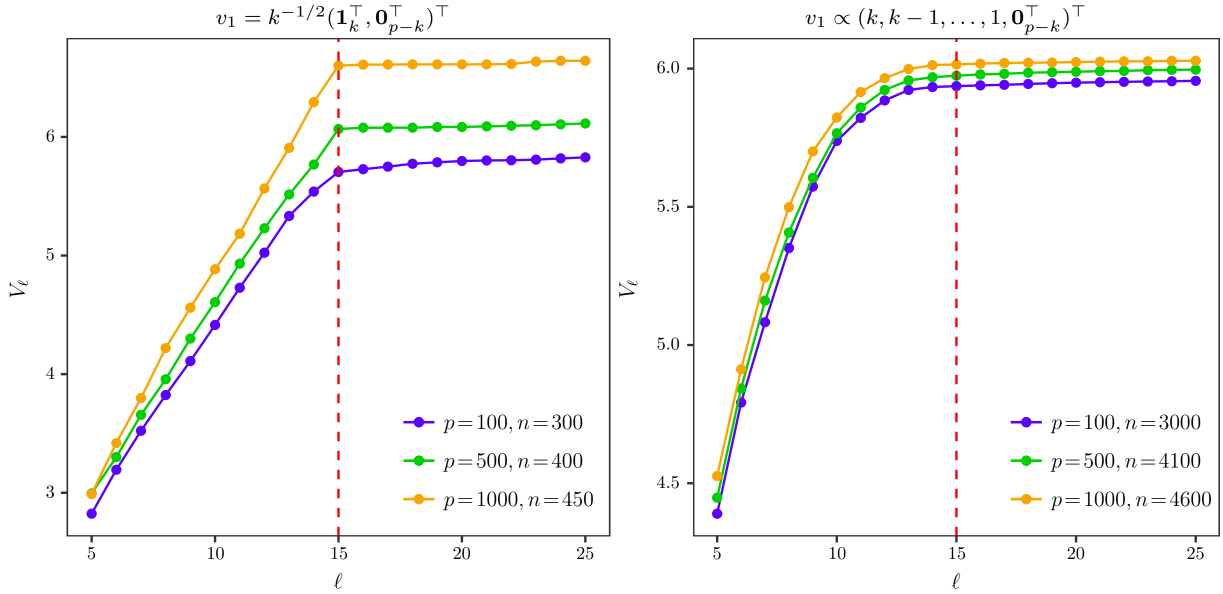


Figure 8: Selecting ℓ. Left: $\theta_1 = 5$, $\theta_2 = 3$, $k = 15$, $d = 10$, $A = 300$, $B = 100$. Right: $\theta_1 = 5$, $k = 15$, $d = 20$, $A = 500$, $B = 100$.

$k = 10$, where Σ takes one of the two following forms:

$$\Sigma_{(1)} = I_{100} + \begin{pmatrix} 0.2J_{10} & & \\ & 0.1J_{10} & \\ & & \mathbf{0} \end{pmatrix}, \qquad \Sigma_{(2)} = \begin{pmatrix} J_{10} & \\ & I_{90} + 0.2J_{90} \end{pmatrix}. \tag{13}$$

Observe that in both examples $v_1 = k^{-1/2}(\mathbf{1}_k^\top, \mathbf{0}_{p-k}^\top)^\top$. Also, the covariance matrix $\Sigma_{(1)}$ is double-spiked with $\theta_1 = 2$, $\theta_2 = 1$ and $v_2 = k^{-1/2}(\mathbf{0}_k^\top, \mathbf{1}_k^\top, \mathbf{0}_{p-2k}^\top)^\top$. We compare the empirical performance of our algorithm with methods proposed by Zou, Hastie and Tibshirani (2006); Shen and Huang (2008); d'Aspremont, Bach and El Ghaoui (2008); Witten, Tibshirani and Hastie (2009) and Ma (2013), by computing the average loss for each algorithm over 100 repetitions on the same set of data. We note that these are all iterative methods, whose success depends on good initialization. In particular, the methods by Zou, Hastie and Tibshirani (2006); Shen and Huang (2008) and Witten, Tibshirani and Hastie (2009) use singular value decomposition of the sample covariance matrix to compute their initial point, while the methods by d'Aspremont, Bach and El Ghaoui (2008) and Ma (2013) select their initialization according to largest diagonal entries of $\hat{\Sigma}$. We have not included estimation procedures based on semidefinite programming (e.g d'Aspremont et al., 2007) as they were computationally very expensive and not as competitive in our simulation studies.

In Figure 9, we see that for the example with spiked covariance $\Sigma_{(1)}$, the performance of the SPCAvRP estimator compares well with other approaches, though the method of Shen and Huang (2008) with hard-thresholding (ℓ_0-thresholding) performs slightly better for small n. However, similarly as in the example from the introduction, if the covariance matrix is given by $\Sigma_{(2)}$, only SPCAvRP produces a consistent estimator among the tested algorithms. All other methods initialize at a vector whose support is disjoint from the true signal coordinates and this feature remains throughout their iterative updates.

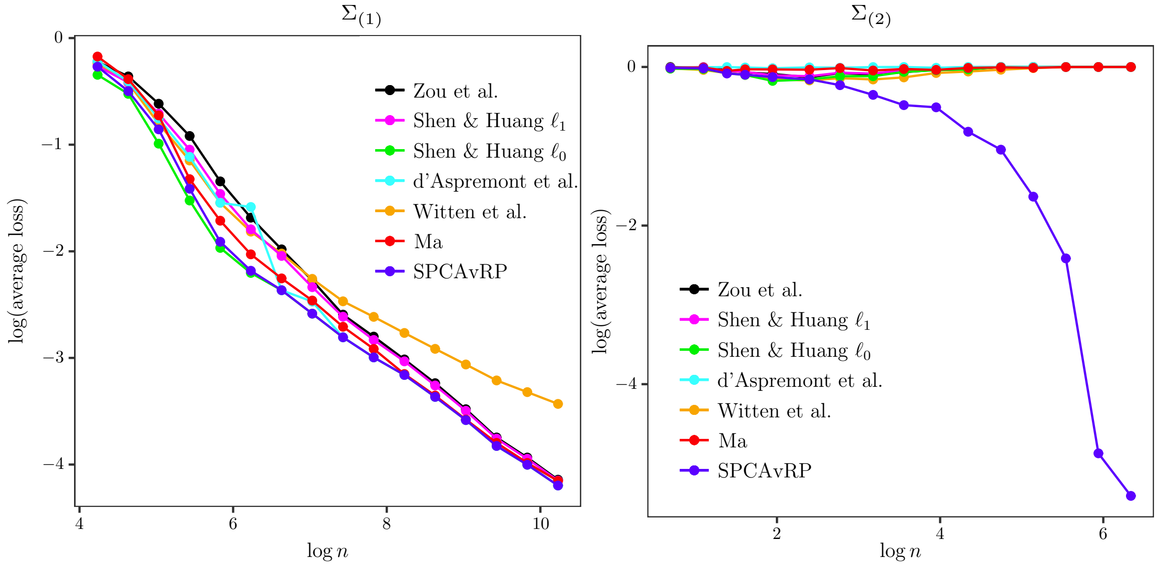


Figure 9: Loss against sample size n using two different covariance structures from (13). Blue: SPCAvRP with $A = 300, B = 150$, $d = l = k$; black: Zou, Hastie and Tibshirani (2006) with given k; magenta and green: Shen and Huang (2008) with ℓ_1 and ℓ_0-thresholding, respectively, both with given k; cyan: d'Aspremont, Bach and El Ghaoui (2008) with given k; orange: Witten, Tibshirani and Hastie (2009) with parameters chosen by their cross-validation; red: Ma (2013) with the default parameters.

4.3.2 Higher-order components

In Table 1 and Figure 10 we compare Algorithm 3 with existing SPCA algorithms for subspace estimation, namely those proposed by Zou, Hastie and Tibshirani (2006), Witten, Tibshirani and Hastie (2009) and Ma (2013). For this purpose we simulate observations from a normal distribution with a covariance matrix which is two- and three-spiked, respectively. Besides computing the loss function for individual components $L(\hat{v}_r, v_r)$, we also compute the error incurred in estimating the subspace spanned by $v_1, \ldots, v_s$ as $\|P_{\hat{V}_s} - P_{V_s}\|_{\mathrm{op}}$, where $P_{\hat{V}_s}$ and P_{V_s} denote the orthogonal projections onto $\mathrm{span}\{\hat{v}_1, \ldots, \hat{v}_s\}$ and $\mathrm{span}\{v_1, \ldots, v_s\}$ respectively. Each evaluation of the loss function is averaged over 100 repetitions. From Table 1 and Figure 10, we observe that our SPCAvRP performs very well with respect to both types of loss function when compared with the alternative algorithms. From Table 1, we also see that only our algorithm and the one proposed by Ma (2013) compute components that are orthogonal in both cases $S_1 \cap S_2 = \emptyset$ and $S_1 \cap S_2 \neq \emptyset$, where $S_r := \{j \in [p] : v_r^{(j)} \neq 0\}$. Although in Figure 10 only the case $S_1 \cap S_2 \cap S_3 \neq \emptyset$ is presented, we observed similar performance when the supports are disjoint.

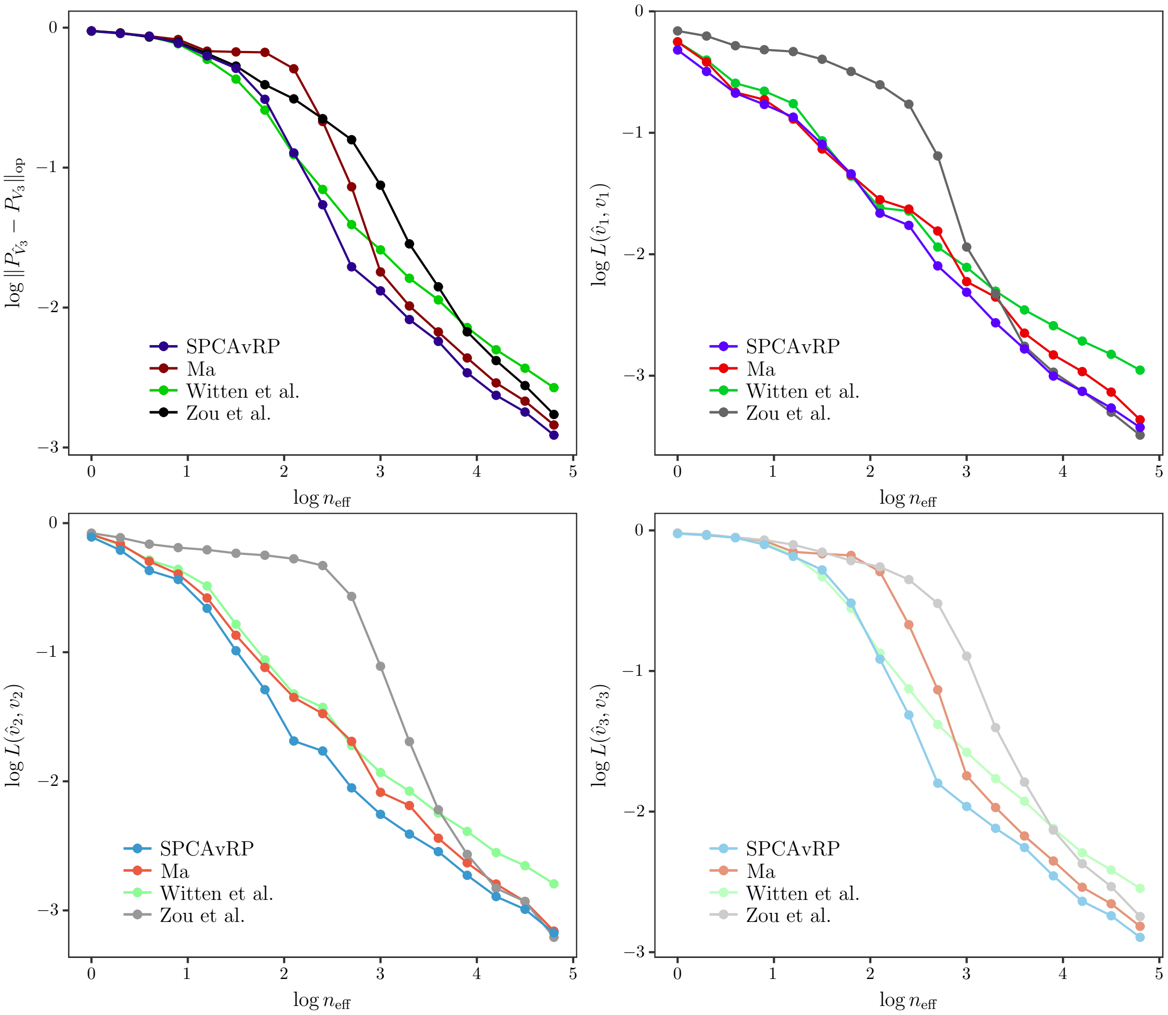


Figure 10: Observations are generated from $N_p(0, \Sigma)$, $\Sigma = I_p + \theta_1 v_1 v_1^\top + \theta_2 v_2 v_2^\top + \theta_3 v_3 v_3^\top$, $\theta_1 = 3$, $\theta_2 = 2$, $\theta_3 = 1$, $p = 100$, where v_1, v_2, v_3 have homogeneous signals strengths with $S_1 = \{1, \ldots, 10\}$, $S_2 = \{5, \ldots, 14\}$ and $S_3 = \{9, \ldots, 18\}$. The four figures correspond to loss functions $\|P_{\hat{V}_3} - P_{V_3}\|_{\mathrm{op}}$, $L(\hat{v}_1, v_1)$, $L(\hat{v}_2, v_2)$ and $L(\hat{v}_3, v_3)$. SPCAvRP algorithm given in Algorithm 3 with input $A = 400$, $B = 200$, $s = 3$, $d = \ell_1 = \ell_2 = \ell_3 = k$, is compared with algorithms for subspace estimation proposed by Zou, Hastie and Tibshirani (2006), Witten, Tibshirani and Hastie (2009) and Ma (2013), which are used with their default parameters.

$S_1 \cap S_2 = \emptyset$	$\|P_{\hat{V}_2} - P_{V_2}\|_{\mathrm{op}}$	$L(\hat{v}_1, v_1)$	$L(\hat{v}_2, v_2)$	$\lvert\hat{v}_1^\top \hat{v}_2\rvert$
SPCAvRP	5.69×10^{-2}	4.50×10^{-2}	5.49×10^{-2}	$< 10^{-15}$
Ma	8.47×10^{-2}	1.30×10^{-1}	1.37×10^{-1}	$< 10^{-15}$
Witten et al.	8.13×10^{-2}	9.75×10^{-2}	9.60×10^{-2}	4.61×10^{-3}
Zou et al.	9.60×10^{-2}	7.11×10^{-2}	9.39×10^{-2}	$< 10^{-15}$

$S_1 \cap S_2 \neq \emptyset$	$\|P_{\hat{V}_2} - P_{V_2}\|_{\mathrm{op}}$	$L(\hat{v}_1, v_1)$	$L(\hat{v}_2, v_2)$	$\lvert\hat{v}_1^\top \hat{v}_2\rvert$
SPCAvRP	7.64×10^{-2}	7.97×10^{-2}	8.39×10^{-2}	$< 10^{-15}$
Ma	7.85×10^{-2}	1.31×10^{-1}	1.42×10^{-1}	$< 10^{-15}$
Witten et al.	9.20×10^{-2}	1.31×10^{-1}	1.33×10^{-1}	9.04×10^{-4}
Zou et al.	1.63×10^{-1}	1.84×10^{-1}	2.00×10^{-1}	5.84×10^{-4}

Table 1: Observations are generated from $N_p(0, \Sigma)$, $\Sigma = I_p + \theta_1 v_1 v_1^\top + \theta_2 v_2 v_2^\top$, $\theta_1 = 50$, $\theta_2 = 30$, $p = 200$, $n = 150$, where v_1 and v_2 have homogeneous signal strengths with $S_1 = \{1, \ldots, 14\}$, and $S_2 = \{15, \ldots, 28\}$ (top), $S_2 = \{7, \ldots, 20\}$ (bottom). The SPCAvRP algorithm given in Algorithm 3 with input $A = 300$, $B = 150$, $s = 2$, $d = \ell_1 = \ell_2 = k$, is compared with algorithms proposed by Zou, Hastie and Tibshirani (2006), Witten, Tibshirani and Hastie (2009) and Ma (2013), which are used with their default parameters.

4.3.3 Microarray data

We test our SPCAvRP algorithm on the Alon et al. (1999) gene expression data set, which contains 40 colon tumor and 22 normal observations. A preprocessed data set can be downloaded from the R package 'datamicroarray' Ramey (2016), with a total of $p = 2000$ features and $n = 62$ observations. For comparison with alternative SPCA approaches, we use algorithms that accept the output sparsity ℓ as an input parameter, namely those proposed by Zou, Hastie and Tibshirani (2006), d'Aspremont, Bach and El Ghaoui (2008) and Shen and Huang (2008). For each ℓ considered, we computed the estimator $\hat{v}_{1,\ell}$ of the first principal component, and in Figure 11 we plot the explained variance $V_\ell := \hat{v}_{1,\ell}^\top \hat{\Sigma} \hat{v}_{1,\ell}$ as well as two different metrics for the separability of the two classes of observations projected along first principal component $\hat{v}_{1,\ell}$, namely the Wasserstein distance W_ℓ of order one and the p-value of Welch's t-test (Welch, 1947). Furthermore, in Figure 12, we display their corresponding values for $\ell = 20$ together with the box plots of the observations from the two classes projected along $\hat{v}_{1,20}$. From Figures 11 and 12, we observe that the SPCAvRP algorithm performs similarly to those proposed by d'Aspremont, Bach and El Ghaoui (2008) and Shen and Huang (2008), all of which are superior in this instance to the SPCA algorithm of Zou, Hastie and Tibshirani (2006). In particular, for small values of ℓ, we observe a steep slope of the blue Wasserstein and p-value curves corresponding to SPCAvRP algorithm in Figure 11, indicating that the two classes are well separated by projecting the observations along the estimated principal component which contains expression levels of only a few different genes.

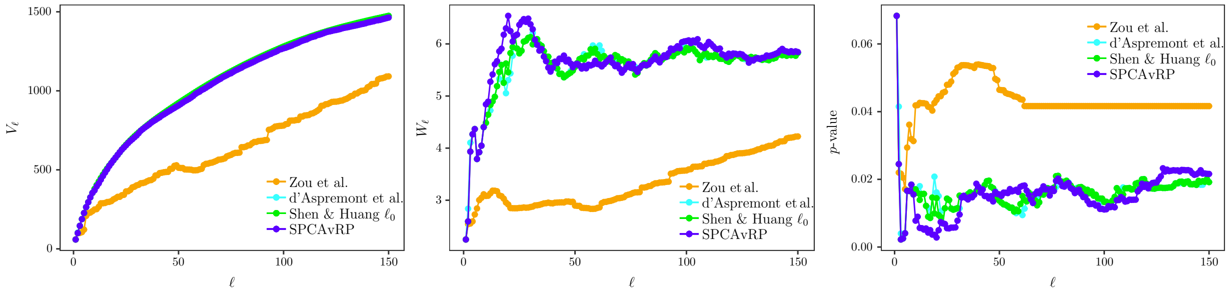
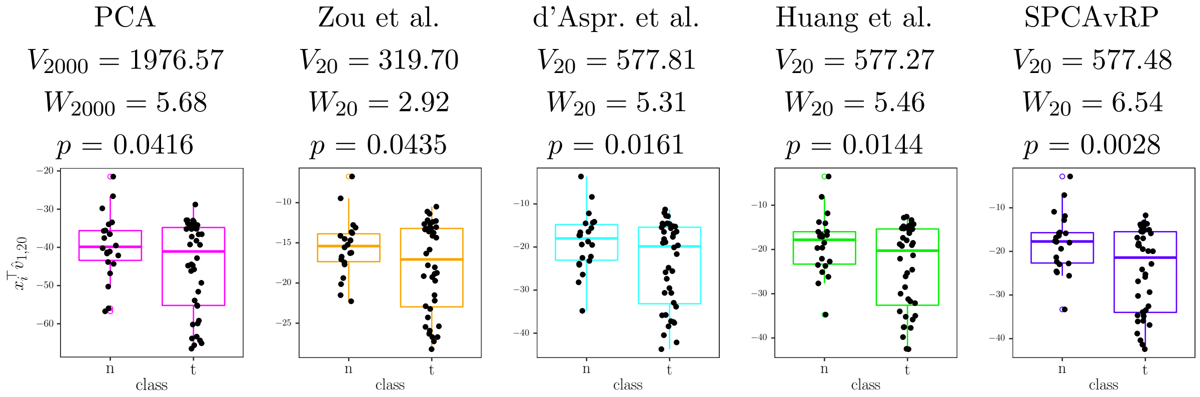


Figure 11: Left panel: V_ℓ; middle panel: Wasserstein distance W_ℓ between the empirical distributions of the two classes projected along $\hat{v}_{1,\ell}$; right panel: p-value of Welch's t-test for the two classes projected along $\hat{v}_{1,\ell}$, where $\hat{v}_{1,\ell}$ is the estimator of v_1 for varied sparsity level ℓ. For estimation we use SPCAvRP ($d = 30$, $A = 1200$, $B = 200$), Zou, Hastie and Tibshirani (2006), d'Aspremont, Bach and El Ghaoui (2008) and Shen and Huang (2008) with ℓ_0-thresholding.


PCA
$V_{2000} = 1976.57$
$W_{2000} = 5.68$
$p = 0.0416$

Zou et al.
$V_{20} = 319.70$
$W_{20} = 2.92$
$p = 0.0435$

d'Aspr. et al.
$V_{20} = 577.81$
$W_{20} = 5.31$
$p = 0.0161$

Huang et al.
$V_{20} = 577.27$
$W_{20} = 5.46$
$p = 0.0144$

SPCAvRP
$V_{20} = 577.48$
$W_{20} = 6.54$
$p = 0.0028$

Figure 12: Variance V_ℓ, Wasserstein distance W_ℓ, p-value of the Welch's t-test and the corresponding box plots of the observations from the two classes projected along estimator $\hat{v}_{1,\ell}$ of the first principal component computed by five different approaches: classical PCA, Zou, Hastie and Tibshirani (2006), d'Aspremont, Bach and El Ghaoui (2008), Shen and Huang (2008) with ℓ_0-thresholding, and SPCAvRP . The desired sparsity level in all SPCA algorithms is set to $\ell = 20$.

A Proofs of theoretical results

Proof of Lemma 1. To verify that $\hat{v}_r$ is orthogonal to $\hat{v}_1, \ldots, \hat{v}_{r-1}$, observe that since the support of $\hat{v}_r$ is contained in $\tilde{S}_r$, we have

$$\hat{v}_r^\top \hat{V}_{r-1} = \hat{v}_r^\top P_{\tilde{S}_r} \hat{V}_{r-1} + \hat{v}_r^\top P_{\tilde{S}_r^c} \hat{V}_{r-1} = \frac{\hat{v}_r^\top H_{\tilde{S}_r} P_{\tilde{S}_r} \hat{\Sigma} P_{\tilde{S}_r} H_{\tilde{S}_r} P_{\tilde{S}_r} \hat{V}_{r-1}}{\lambda_1(H_{\tilde{S}_r} P_{\tilde{S}_r} \hat{\Sigma} P_{\tilde{S}_r} H_{\tilde{S}_r})} = 0,$$

where the final equality follows from the fact that $H_{\tilde{S}_r}$ is a projection onto the orthogonal complement of the column space of $P_{\tilde{S}_r} \hat{V}_{r-1}$, so $H_{\tilde{S}_r} P_{\tilde{S}_r} \hat{V}_{r-1} = 0$. □

Proof of Theorem 2. For each $r \in [p]$, we define an event

$$\Omega_{0,r} := \left\{ \|P_S(\hat{\Sigma} - \Sigma)P_S\|_{\mathrm{op}} \leq 2K\sqrt{\frac{r \log p}{n}} \text{ for all } S \subseteq [p] \text{ and } |S| = r \right\}.$$

Since $Q \in \mathrm{RCC}_p(K)$, by choosing $\delta = p^{-3}$ in (6) and using the fact that $4\max(d, \ell)\log p \leq n$, we have $\mathbb{P}(\Omega_{0,d}) \geq 1 - p^{-3}$ and $\mathbb{P}(\Omega_{0,\ell}) \geq 1 - p^{-3}$.

For each $a \in [A]$ and $b \in [B]$, let $S_{a,b}$ denote the set of indices corresponding to ones on the diagonal of projection $P_{a,b}$. If $s_{a,b} := |S_{a,b} \cap [k]| \geq 1$, then we have $\lambda_{a,b} := \lambda_1(P_{a,b}\Sigma P_{a,b}) = 1 + s_{a,b}\theta_1/k$ and $v_{a,b} := v_1(P_{a,b}\Sigma P_{a,b}) = s_{a,b}^{-1/2}(\mathbb{1}_{\{j \in S_{a,b} \cap [k]\}})_{j \in [p]}$ (up to sign). We similarly define $\hat{\lambda}_{a,b} := \lambda_1(P_{a,b}\hat{\Sigma}P_{a,b})$ and $\hat{v}_{a,b} := v_1(P_{a,b}\hat{\Sigma}P_{a,b})$ (with signs chosen such that $v_{a,b}^\top \hat{v}_{a,b} \geq 0$). Recall that by definition $b^*(a) = \mathrm{sargmax}_{b \in [B]} \hat{\lambda}_{a,b}$. On $\Omega_{0,d}$, by Weyl's inequality and (8), we have that

$$\max_{a \in [A], b \in [B]} \bigl|\hat{\lambda}_{a,b} - \lambda_{a,b}\bigr| \leq \max_{a \in [A], b \in [B]} \|P_{S_{a,b}}(\hat{\Sigma} - \Sigma)P_{S_{a,b}}\|_{\mathrm{op}} \leq 2K\sqrt{\frac{d \log p}{n}} \leq \frac{t\theta_1}{20k}.$$

Thus, if $\bar{b}(a) \in \mathrm{argmax}_{b \in [B]} s_{a,b}$, then on $\Omega_{0,d}$, we have

$$\begin{aligned} s_{a,b^*(a)} = \frac{k}{\theta_1}(\lambda_{a,b^*(a)} - 1) \geq \frac{k}{\theta_1}(\hat{\lambda}_{a,b^*(a)} - 1) - \frac{t}{20} &\geq \frac{k}{\theta_1}(\hat{\lambda}_{a,\bar{b}(a)} - 1) - \frac{t}{20} \\ &\geq \frac{k}{\theta_1}(\lambda_{a,\bar{b}(a)} - 1) - \frac{t}{10} = s_{a,\bar{b}(a)} - \frac{t}{10}. \end{aligned}$$

For each $a \in [A]$, define $\Omega_a := \Omega_{0,d} \cap \{s_{a,\bar{b}(a)} \geq t\}$. Then by (7), we have

$$\mathbb{P}(\Omega_a^{\mathrm{c}}) \leq p^{-3} + F_{\mathrm{HG}}(t-1; d, k, p)^B \leq p^{-3} + \left(1 - \frac{3}{B}\log p\right)^B \leq 2p^{-3}.$$

The first and second eigenvalues of $P_{a,b^*(a)}\Sigma P_{a,b^*(a)}$ are separated by $s_{a,b^*(a)}\theta_1/k$. Hence, by a variant of the Davis–Kahan theorem (Yu, Wang and Samworth, 2015, Corollary 1) (restated for convenience as Lemma A.1 below), on Ω_a we have

$$\bigl\|\hat{v}_{a,b^*(a)} - v_{a,b^*(a)}\bigr\| \leq 2^{5/2}K\sqrt{\frac{dk^2 \log p}{n\theta_1^2 s_{a,b^*(a)}^2}} \leq 7K\sqrt{\frac{dk^2 \log p}{t^2 n\theta_1^2}}.$$

By the triangle inequality, we obtain that

$$\sum_{j=1}^{k}\bigl|\hat{v}^{(j)}_{a,b^*(a)}\bigr| \geq \biggl\{\sum_{j=1}^{k}\bigl(\hat{v}^{(j)}_{a,b^*(a)}\bigr)^2\biggr\}^{1/2} \geq \biggl(1 - 7K\sqrt{\frac{dk^2\log p}{t^2n\theta_1^2}}\biggr)\mathbb{1}_{\Omega_a},$$

and

$$\sum_{j=k+1}^{p}\bigl|\hat{v}^{(j)}_{a,b^*(a)}\bigr| \leq d^{1/2}\biggl\{\sum_{j\in S_{a,b^*(a)}\cap\{k+1,\ldots,p\}}\bigl(\hat{v}^{(j)}_{a,b^*(a)}\bigr)^2\biggr\}^{1/2} \leq 7K\sqrt{\frac{d^2k^2\log p}{t^2n\theta_1^2}}\mathbb{1}_{\Omega_a} + d^{1/2}\mathbb{1}_{\Omega_a^{\mathrm{c}}}.$$

By symmetry of Q in coordinates $\{1,\ldots,k\}$ and coordinates $\{k+1,\ldots,p\}$ and the fact that $p \geq 4$, we have that for $j \in [k]$ and any $a \in [A]$,

$$\mathbb{E}\bigl|\hat{v}^{(j)}_{a,b^*(a)}\bigr| \geq \biggl(\frac{1}{k} - 7K\sqrt{\frac{d\log p}{t^2n\theta_1^2}}\biggr)\mathbb{P}(\Omega_a) \geq \frac{33}{40k}(1-2p^{-3}) > \frac{3}{4k}. \tag{14}$$

Moreover, since we also have $p \geq 2k$, we deduce that for $j \in \{k+1,\ldots,p\}$ and any $a \in [A]$,

$$\mathbb{E}\bigl|\hat{v}^{(j)}_{a,b^*(a)}\bigr| \leq \frac{7K}{p-k}\sqrt{\frac{d^2k^2\log p}{t^2n\theta_1^2}} + \frac{d^{1/2}}{p-k}\mathbb{P}(\Omega_a^{\mathrm{c}}) \leq \frac{7}{40k} + \frac{2d^{1/2}}{(p-k)p^3} < \frac{1}{4k}, \tag{15}$$

where the first inequality uses Assumption (8) and the final inequality uses the fact that $p \geq 2k$.

Recall the definitions of $\hat{w}^{(j)}$ and $\hat{S}_1$ from Algorithm 1. Let $\Omega := \{\min_{j\leq k}\hat{w}^{(j)} > \max_{j>k}\hat{w}^{(j)}\}$. By (14), (15), a union bound and Hoeffding's inequality, we have that

$$\mathbb{P}(\Omega^{\mathrm{c}}) \leq \mathbb{P}\biggl(\exists\, j\in[p] \text{ s.t. } |\hat{w}^{(j)} - \mathbb{E}\hat{w}^{(j)}| > \frac{1}{4k}\biggr) \leq pe^{-A/(32k^2)}.$$

Define $\tilde{v}_1 \in \mathbb{R}^p$ by $\tilde{v}_1 = v_1(P_{\hat{S}_1}\Sigma P_{\hat{S}_1})$. On Ω, we have $\hat{S}_1 \subseteq [k] = S_1$ if $\ell \leq k$ and $\hat{S}_1 \supseteq [k]$ if $\ell \geq k$. Thus, on the event $\Omega\cap\Omega_{0,\ell}$, which has probability at least $1 - p^{-3} - pe^{-A/(32k^2)}$, we have by Lemma A.1 that

$$\begin{aligned} L(\hat{v}_1, v_1) \leq L(\hat{v}_1,\tilde{v}_1) + L(\tilde{v}_1, v_1) &\leq \frac{2\|P_{\hat{S}_1}(\hat{\Sigma}-\Sigma)P_{\hat{S}_1}\|_{\mathrm{op}}}{\min(\ell,k)\theta_1/k} + \sqrt{\frac{k-\min(\ell,k)}{k}} \\ &\leq 4K\sqrt{\frac{\ell\log p}{n\theta_1^2}}\max\Bigl(1,\frac{k}{\ell}\Bigr) + \sqrt{\max\Bigl(1-\frac{\ell}{k},0\Bigr)}, \end{aligned}$$

as desired. □

The following lemma, which is used several times in our proofs, is a special case of the variant of the Davis–Kahan theorem given in Yu, Wang and Samworth (2015, Corollary 1).

Lemma A.1. *Let $p \geq 2$, and let $\Sigma, \hat{\Sigma} \in \mathbb{R}^{p \times p}$ be symmetric, with eigenvalues $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_p$ and $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$ respectively. If $v_1, \hat{v}_1 \in \mathbb{R}^p$ satisfy $\Sigma v_1 = \lambda_1 v_1$ and $\hat{\Sigma}\hat{v}_1 = \hat{\lambda}_1 \hat{v}_1$, then*

$$L(\hat{v}_1, v_1) \leq \frac{2\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}}{\lambda_1 - \lambda_2}.$$

Moreover, if $\hat{v}_1^\top v_1 \geq 0$, then

$$\|\hat{v}_1 - v_1\| \leq \frac{2^{3/2}\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}}{\lambda_1 - \lambda_2}.$$

Proof of Proposition 3. We observe that

$$\begin{aligned} L(\hat{v}_1, v_1) &\leq L(\hat{v}_1, v_1)\mathbb{1}_{\{S_1 \subseteq \hat{S}_1\}} + \mathbb{1}_{\{S_1 \not\subseteq \hat{S}_1\}} \\ &\leq \frac{2}{\lambda_1 - \lambda_2}\|P_{\hat{S}_1}\hat{\Sigma}P_{\hat{S}_1} - P_{\hat{S}_1}\Sigma P_{\hat{S}_1}\|_{\mathrm{op}}\mathbb{1}_{\{S_1 \subseteq \hat{S}_1\}} + \mathbb{1}_{\{S_1 \not\subseteq \hat{S}_1\}}, \end{aligned} \tag{16}$$

where the second inequality follows from Lemma A.1 and the fact that for any S such that $S_1 \subseteq S$ we have $v_1 = v_1(P_S \Sigma P_S)$, $\lambda_1 = \lambda_1(P_S \Sigma P_S)$, and also $\lambda_1(P_S \Sigma P_S) - \lambda_2(P_S \Sigma P_S) \geq \lambda_1 - \lambda_2 > 0$. Next, for the first term in (16), we have

$$\|P_{\hat{S}_1}\hat{\Sigma}P_{\hat{S}_1} - P_{\hat{S}_1}\Sigma P_{\hat{S}_1}\|_{\mathrm{op}} = \sup_{v \in \mathcal{S}^{p-1}} v^\top P_{\hat{S}_1}(\hat{\Sigma} - \Sigma)P_{\hat{S}_1}v \leq \sup_{v \in \mathcal{B}_0^{p-1}(\ell)} \bigl|v^\top(\hat{\Sigma} - \Sigma)v\bigr|.$$

Thus, by the definition of the RCC condition in (6) and Wang, Berthet and Samworth (2016b, Proposition 1), after taking expectation on both sides of (16) we obtain the desired bound. □

The following two lemmas are used in the proof of Theorem 4. The first shows the implications of having a projection P in the set of good projections $\mathcal{G}$.

Lemma A.2. *Let $P \in \mathcal{P}_d$ and let $\Sigma \in \mathbb{R}^{p \times p}$ be positive semidefinite with leading eigenvalues $\lambda_1 \geq \lambda_2$. If $\|Pv_1\| \geq \tau > 0$, then $L\bigl(v_1(P\Sigma P), Pv_1/\|Pv_1\|\bigr) \leq 2\lambda_2/(\lambda_1\tau^2)$ and $\lambda_1(P\Sigma P) - \lambda_2(P\Sigma P) \geq \lambda_1\tau^2 - \lambda_2$.*

Remark: In the case where Σ is a spiked covariance matrix of the form (12), the conclusions of the lemma are that $L\bigl(v_1(P\Sigma P), Pv_1/\|Pv_1\|\bigr) \leq 2\theta_2/(\theta_1\tau^2)$ and $\lambda_1(P\Sigma P) - \lambda_2(P\Sigma P) \geq \theta_1\tau^2 - \theta_2$; cf. the discussion following Theorem 4.

Proof of Lemma A.2. By eigendecomposition, we may write $\Sigma = \sum_{r=1}^p \lambda_r v_r v_r^\top$, where $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ and where $v_1, \ldots, v_p \in \mathbb{R}^p$ are orthonormal. Writing $S := \lambda_1 P v_1 v_1^\top P$ and $E := \sum_{r=2}^p \lambda_r P v_r v_r^\top P$, we have $P\Sigma P = S + E$. Observe that $v_1(S) = Pv_1/\|Pv_1\|$ and $\lambda_1(S) - \lambda_2(S) = \lambda_1\|Pv_1\|^2 \geq \lambda_1\tau^2$. On the other hand,

$$\|E\|_{\mathrm{op}} \leq \biggl\|\sum_{r=2}^m \lambda_r v_r v_r^\top\biggr\|_{\mathrm{op}} = \lambda_2. \tag{17}$$

Thus, by Lemma A.1, we have

$$L\biggl(v_1(P\Sigma P), \frac{Pv_1}{\|Pv_1\|}\biggr) \leq \frac{2\|E\|_{\mathrm{op}}}{\lambda_1(S) - \lambda_2(S)} \leq \frac{2\lambda_2}{\lambda_1 \tau^2},$$

which establishes the first claim. For the second claim, by Weyl's inequality (Weyl, 1912; Stewart and Sun, 1990, Corollary IV.4.9), for every $r \in [p]$, we have $\lambda_r(S) + \lambda_p(E) \leq \lambda_r(P\Sigma P) \leq \lambda_r(S) + \lambda_1(E)$. Since E is positive semidefinite, we deduce that

$$\lambda_1(P\Sigma P) - \lambda_2(P\Sigma P) \geq \lambda_1(S) - \lambda_2(S) - \lambda_1(E) \geq \lambda_1 \tau^2 - \lambda_2,$$

as desired. □

The next lemma shows that, when B is sufficiently large, we have $\|P_1 v_1\| \geq \tau > 0$ with high probability, where P_1 is selected from a group of B random projections via our selection criteria (4) in Algorithm 1.

Lemma A.3. *Let $P_1 := P_{1,b^*(1)}$ be selected as in Algorithm 1 with input $X_1, \ldots, X_n$, B and d. Assume further that $X_1, \ldots, X_n \overset{\mathrm{iid}}{\sim} Q \in \mathrm{RCC}_p(K)$ with covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$. If $n \geq 4d \log p$ and $p \geq 3$, then for any $k' \in [k]$ and $\tau > 0$,*

$$\mathbb{P}(\|P_1 v_1\| < \tau) \leq p^{-3} + F_{\mathrm{HG}}\Biggl(\frac{1}{v_{1,(k')}^2}\biggl\{\tau^2 + \frac{4K}{\lambda_1}\sqrt{\frac{d\log p}{n}} + \frac{\lambda_2}{\lambda_1}\biggr\}; d, k', p\Biggr)^B.$$

Remark: Similarly as in Lemma A.2 , in the case where Σ is a spiked covariance matrix of the form (12), λ_r in the claim of this lemma can be replaced by θ_r, $r = 1, 2$.

Proof of Lemma A.3. Taking $\delta = p^{-3}$ in (6), there is an event Ω with probability at least $1 - p^{-3}$ on which for all $b \in [B]$,

$$\bigl|\lambda_1\bigl(P_{1,b}\hat{\Sigma}P_{1,b}\bigr) - \lambda_1\bigl(P_{1,b}\Sigma P_{1,b}\bigr)\bigr| \leq \bigl\|P_{1,b}\hat{\Sigma}P_{1,b} - P_{1,b}\Sigma P_{1,b}\bigr\|_{\mathrm{op}} \leq 2K\sqrt{\frac{d\log p}{n}}. \tag{18}$$

By Weyl's inequality as in the proof of Lemma A.2, we have for any $b \in [B]$ that

$$\lambda_1\|P_{1,b}v_1\|^2 \leq \lambda_1(P_{1,b}\Sigma P_{1,b}) \leq \lambda_1\|P_{1,b}v_1\|^2 + \lambda_2. \tag{19}$$

Recall that $P_1 = P_{1,b^*}$, where $b^* = \mathrm{sargmax}_{b\in[B]}\,\lambda_1\bigl(P_{1,b}\hat{\Sigma}P_{1,b}\bigr)$. In addition, we define $\tilde{b} := \mathrm{sargmax}_{b\in[B]}\,\|P_{1,b}v_1\|$. Then by (18) and (19), we have on Ω that

$$\begin{aligned}\lambda_1\|P_1v_1\|^2 &\geq \lambda_1(P_1\hat{\Sigma}P_1) - 2K\sqrt{\frac{d\log p}{n}} - \lambda_2 \geq \lambda_1(P_{1,\tilde{b}}\hat{\Sigma}P_{1,\tilde{b}}) - 2K\sqrt{\frac{d\log p}{n}} - \lambda_2 \\ &\geq \lambda_1(P_{1,\tilde{b}}\Sigma P_{1,\tilde{b}}) - 4K\sqrt{\frac{d\log p}{n}} - \lambda_2 \geq \lambda_1\|P_{1,\tilde{b}}v_1\|^2 - 4K\sqrt{\frac{d\log p}{n}} - \lambda_2.\end{aligned}$$

Consequently, writing $R := \frac{4K}{\lambda_1}\sqrt{\frac{d\log p}{n}} + \frac{\lambda_2}{\lambda_1}$, we have for any $k' \in [k]$ that

$$\begin{aligned}\mathbb{P}\bigl(\|P_1 v_1\| < \tau\bigr) &\le \mathbb{P}(\Omega^{\mathrm{c}}) + \mathbb{P}\bigl(\|P_{1,\tilde{b}} v_1\|^2 < \tau^2 + R\bigr) \\ &\le p^{-3} + \prod_{b\in[B]} \mathbb{P}\bigl(\|P_{1,b} v_1\|^2 < \tau^2 + R\bigr) \le p^{-3} + F_{\mathrm{HG}}\bigl(v_{1,(k')}^{-2}(\tau^2 + R); d, k', p\bigr)^B,\end{aligned}$$

as desired. □

Proof of Theorem 4. Recall that $P_1 = P_{1,b^*(1)}$ and that $\Omega = \{P_1 \in \mathcal{G}\}$. By Lemma A.3 and (11), we have that

$$\mathbb{P}(\Omega^{\mathrm{c}}) \le p^{-3} + \exp\biggl[-B\biggl\{1 - F_{\mathrm{HG}}\biggl(\frac{1}{v_{1,(k')}^2}\biggl(\tau^2 + \frac{4K}{\lambda_1}\sqrt{\frac{d\log p}{n}} + \frac{\lambda_2}{\lambda_1}\biggr); d, k', p\biggr)\biggr\}\biggr] \le 2p^{-3}. \tag{20}$$

Moreover, on Ω, we have by Lemma A.2 that $L\bigl(v_1(P\Sigma P), Pv_1/\|Pv_1\|\bigr) \le 2\lambda_2/(\lambda_1\tau^2) =: \mu$ and $\lambda_1(P\Sigma P) - \lambda_2(P\Sigma P) \ge \lambda_1\tau^2 - \lambda_2 =: \kappa$. Now let

$$\xi := \frac{4\sqrt{2}K}{\kappa}\sqrt{\frac{d\log p}{n}} + \sqrt{2}\mu$$

and $u_1 := P_1 v_1/\|P_1 v_1\|$. Since $d\log p \le n$, by Lemma A.1 and Wang, Berthet and Samworth (2016b, Proposition 1), we have

$$\begin{aligned}\mathbb{E}\bigl(L(\hat{v}_{1,b^*(1)}, u_1)\mathbb{1}_\Omega\bigr) &\le \mathbb{E}\bigl(L(\hat{v}_{1,b^*(1)}, v_1(P_1\Sigma P_1))\mathbb{1}_\Omega\bigr) + \mathbb{E}\bigl(L(v_1(P_1\Sigma P_1), u_1)\mathbb{1}_\Omega\bigr) \\ &\le \frac{2}{\kappa}\,\mathbb{E}\bigl(\|P_1\hat{\Sigma}P_1 - P_1\Sigma P_1\|_{\mathrm{op}}\mathbb{1}_\Omega\bigr) + \mu \le \frac{\xi}{\sqrt{2}}.\end{aligned}$$

Thus, using the fact that $\|u - \mathrm{sgn}(u^\top v)v\| \le \sqrt{2}L(u,v)$ for $u, v \in \mathcal{S}^{p-1}$, we have

$$\bigl|\mathbb{E}\bigl(|\hat{v}_{1,b^*(1)}^{(j)}|\mathbb{1}_\Omega\bigr) - \mathbb{E}\bigl(|u_1^{(j)}|\mathbb{1}_\Omega\bigr)\bigr| \le \mathbb{E}\bigl\{\bigl||\hat{v}_{1,b^*(1)}^{(j)}| - |u_1^{(j)}|\bigr|\mathbb{1}_\Omega\bigr\} \le \sqrt{2}\mathbb{E}\bigl\{L(\hat{v}_{1,b^*(1)}, u_1)\mathbb{1}_\Omega\bigr\} \le \xi. \tag{21}$$

Write $\Omega^{(j)} := \{P_1^{(j,j)} = 1\}$, for $j \in [p]$. Then for $j \in S_1$, we have that

$$\begin{aligned}\mathbb{E}|\hat{v}_{1,b^*(1)}^{(j)}| &\ge \mathbb{E}\bigl(|u_1^{(j)}|\mathbb{1}_\Omega\bigr) - \xi \ge |v_1^{(j)}|\mathbb{P}(\Omega\cap\Omega^{(j)}) - \xi \\ &\ge v_1^{\min}\Bigl\{\max_{j'\in S_1}\mathbb{P}(\Omega\cap\Omega^{(j')}) - \rho\Bigr\} - \xi \ge v_1^{\min}\biggl\{\tau^2\biggl(1 - \frac{2}{p^3}\biggr) - \rho\biggr\} - \xi,\end{aligned} \tag{22}$$

where we used (21), the definition of ρ in (9), the fact that

$$\tau^2\mathbb{P}(\Omega) \le \mathbb{E}\bigl(\|P_1 v_1\|^2\mathbb{1}_\Omega\bigr) = \sum_{j'\in S_1}(v_1^{(j')})^2\mathbb{P}(\Omega\cap\Omega^{(j')}) \le \max_{j'\in S_1}\mathbb{P}(\Omega\cap\Omega^{(j')})$$

and (20). For $j \notin S_1$, we have by (21) and (20) that

$$\mathbb{E}|\hat{v}_{1,b^*(1)}^{(j)}| \le \mathbb{E}\bigl(|\hat{v}_{1,b^*(1)}^{(j)}|\mathbb{1}_\Omega\bigr) + \mathbb{P}(\Omega^{\mathrm{c}}) \le \xi + 2p^{-3}. \tag{23}$$

Recall that $\hat{w}^{(j)} = A^{-1}\sum_{a\in[A]} |\hat{v}^{(j)}_{a,b^*(a)}|$, which, conditional on $X_1,\ldots,X_n$, is a sample average of independent and identically distributed random variables. It follows from (22), (23), a union bound and Hoeffding's inequality for $\ell \geq k$ that

$$\mathbb{P}(S_1 \not\subseteq \hat{S}_1) \leq \mathbb{P}\Bigl(\min_{j\in S_1} \hat{w}^{(j)} \leq \max_{j\notin S_1} \hat{w}^{(j)}\Bigr) \leq \mathbb{P}\biggl(\bigcup_{j=1}^{p}\bigl\{|\hat{w}^{(j)} - \mathbb{E}\hat{w}^{(j)}| \geq \varepsilon/2\bigr\}\biggr) \leq pe^{-A\varepsilon^2/8}.$$

The result therefore follows from Proposition 3. □

Acknowledgements: The research of the first and third authors was supported by an Engineering and Physical Sciences Research Council (EPSRC) grant EP/N014588/1 for the centre for Mathematical and Statistical Analysis of Multimodal Clinical Imaging. The second and third authors were supported by EPSRC Fellowship EP/J017213/1 and EP/P031447/1, and grant RG81761 from the Leverhulme Trust.

References

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J.(1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *PNAS*, **96**, 6745–6750.

Amini, A. A. and Wainwright, M. J. (2009) High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Statist.*, **37**, 2877–2921.

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

Cannings, T. I. and Samworth, R. J. (2017) Random-projection ensemble classification. *J. Roy. Statist. Soc., Ser. B (with discussion)*, **79**, 959–1035.

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

d'Aspremont, A., El Ghaoui, L., Jordan, M., I., Lanckriet, G., R., G. (2007) A direct formulation for sparse PCA using semidefinite programming. *Adv. Neural. Inf. Process. Syst.*. **16**, 41–48.

Fowler, J. E. (2009) Compressive-projection principal component analysis. *IEEE Trans. Image Process.*, **18**, 2230–2242.

Gataric, M., Wang, T. and Samworth, R. J. (2017) SPCAvRP: Sparse Principal Component Analysis via Random Projections. R package, available at `https://cran.r-project.org/web/packages/SPCAvRP/index.html`.

Hoeffding, W. (1963) Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, **58**, 13–30.

Johnstone, I. M. and Lu, A. Y. (2009) On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.*, **104**, 682–693.

Jolliffe, I. T., Trendafilov, N. T. and Uddin, M. (2003) A modified principal component technique based on the LASSO. *J. Comput. Graph. Statist.*, **12**, 531–547.

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

Mackey, L. W. (2009) Deflation Methods for Sparse PCA. *Adv. Neural. Inf. Process. Syst.*, **21**, 1017–1024.

Marzetta, T. L., Tucci, G. H. and Simon, S. H. (2011) A random matrix-theoretic approach to handling singular covariance estimates. *IEEE Trans. Inf. Theory*, **57**, 6256–6271.

Moghaddam, B., Weiss, Y., and Avidan, S. (2006) Spectral bounds for sparse PCA: Exact and greedy algorithms. *Adv. Neural. Inf. Process. Syst.*, **18**, 915–922.

Paul, D. (2007) Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica.* **17**, 1617–1642.

Paul, D. and Johnstone, I. M. (2012) Augmented sparse principal component analysis for high dimensional data. *arXiv preprint*, arxiv:1202.1242v1.

Pourkamali-Anaraki, F. and Hughes, S. (2014). Memory and computation efficient PCA via very sparse random projections. *Proceedings of the 31st International Conference on Machine Learning*, 1341–1349.

Qi, H. and Hughes, S. (2012). Invariance of principal components under low-dimensional random projection of the data. *Proceedings of 19th IEEE International Conference on Image Processing*, 937–940.

Ramey, J. A. (2016) Collection of Data Sets for Classification. R package, available at `https://github.com/ramhiser/datamicroarray`.

Shen, H. and Huang, J. Z. (2008) Sparse principal component analysis via regularized low rank matrix approximation. *J. Multivariate Anal.*, **99**, 1015–1034.

Stewart, G. W. and Sun, J.-G. (1990). *Matrix Perturbation Theory.* Academic Press, Inc., San Diego, California.

Tillman, A. N. and Pfetsch, M. E. (2014) The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *IEEE Trans. Inform. Theory*, **60**, 1248–1259.

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

Wang, T., Berthet, Q. and Samworth, R. J. (2016a) Statistical and computational trade-offs in estimation of sparse principal components. *Ann. Statist.*, **44**, 1896–1930.

Wang, T., Berthet, Q. and Samworth, R. J. (2016b) Supplementary material to 'Statistical and computational trade-offs in estimation of sparse principal components'. *Ann. Statist.*.

Welch, B. L. (1947) The generalization of "Student's" problem when several different population variances are involved. *Biometrika*, **34**, 28–35.

Weyl, H. (1912). Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf der Theorie der Hohlraumstrahlung). *Math. Ann.* **71**, 441–479.

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.

Yu, Y., Wang, T. and Samworth, R. J. (2015) A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, **102**, 315–323.

Zou, H., Hastie, T. and Tibshirani, R. (2006) Sparse Principal Components Analysis. *J. Comput. Graph. Statist.*, **15**, 265–286.