Biometrika (2014), **99**, 1, *pp*. 1–9

Printed in Great Britain

Advance Access publication on 31 July 2012

A useful variant of the Davis–Kahan theorem for statisticians

BY Y. YU, T. WANG AND R. J. SAMWORTH

Statistical Laboratory, University of Cambridge
Wilberforce Road
Cambridge CB3 0WB
U.K.

y.yu@statslab.cam.ac.uk t.wang@statslab.cam.ac.uk r.samworth@statslab.cam.ac.uk

SUMMARY

The Davis–Kahan theorem is used in the analysis of many statistical procedures to bound the distance between subspaces spanned by population eigenvectors and their sample versions. It relies on an eigenvalue separation condition between certain relevant population and sample eigenvalues. We present a variant of this result that depends only on a population eigenvalue separation condition, making it more natural and convenient for direct application in statistical contexts, and provide an improvement in many cases to the usual bound in the statistical literature. We also give an extension to situations where the matrices under study may be asymmetric or even non-square, and where interest is in the distance between subspaces spanned by corresponding singular vectors.

Some key words: Davis–Kahan theorem; Eigendecomposition; Matrix perturbation; Singular value decomposition.

1. INTRODUCTION

Many statistical procedures rely on the eigendecomposition of a matrix. Examples include principal components analysis and its cousin sparse principal components analysis (Zou et al., 2006), factor analysis, high-dimensional covariance matrix estimation (Fan et al., 2013) and spectral clustering for community detection with network data (Donath and Hoffman, 1973). In these and most other related statistical applications, the matrix involved is real and symmetric, e.g. a covariance or correlation matrix, or a graph Laplacian or adjacency matrix in the case of spectral clustering.

In the theoretical analysis of such methods, it is frequently desirable to be able to argue that if a sample version of this matrix is close to its population counterpart, and provided certain relevant eigenvalues are well-separated in a sense to be made precise below, then a population eigenvector should be well approximated by a corresponding sample eigenvector. A quantitative version of such a result is provided by the Davis–Kahan $\sin\theta$ theorem (Davis and Kahan, 1970). This is a deep theorem from operator theory, involving operators acting on Hilbert spaces, though as remarked by Stewart and Sun (1990), its 'content more than justifies its impenetrability'. In statistical applications, we typically do not require this full generality; in Theorem 1 below, we state a version in a form typically used in the statistical literature (e.g. von Luxburg, 2007; Rohe et al., 2011). Since the theorem allows for the possibility that more than one eigenvector is of interest, we need to define a notion of distance between subspaces spanned by two sets of vectors. This can be done through the idea of principal angles: if $V, \hat{V} \in \mathbb{R}^{p\times d}$ both have orthonormal columns, then the vector of d principal angles between their column spaces is $(\cos^{-1}\sigma_1, \ldots, \cos^{-1}\sigma_d)^\top$,

where $\sigma_1 \geq \cdots \geq \sigma_d$ are the singular values of $\hat{V}^\top V$. Thus, principal angles between subspaces can be considered as a natural generalization of the acute angle between two vectors. We let $\Theta(\hat{V}, V)$ denote the $d \times d$ diagonal matrix whose jth diagonal entry is the jth principal angle, and let $\sin\Theta(\hat{V}, V)$ be defined entrywise. A convenient way to measure the distance between the column spaces of V and $\hat{V}$ is via $\|\sin\Theta(\hat{V}, V)\|_{\mathrm{F}}$, where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm of a matrix.

THEOREM 1 (DAVIS–KAHAN $\sin\theta$ THEOREM). *Let $\Sigma, \hat{\Sigma} \in \mathbb{R}^{p\times p}$ be symmetric, with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ and $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$ respectively. Fix $1 \leq r \leq s \leq p$, let $d = s - r + 1$, and let $V = (v_r, v_{r+1}, \ldots, v_s) \in \mathbb{R}^{p\times d}$ and $\hat{V} = (\hat{v}_r, \hat{v}_{r+1}, \ldots, \hat{v}_s) \in \mathbb{R}^{p\times d}$ have orthonormal columns satisfying $\Sigma v_j = \lambda_j v_j$ and $\hat{\Sigma}\hat{v}_j = \hat{\lambda}_j \hat{v}_j$ for $j = r, r+1, \ldots, s$. Write $\delta = \inf\{|\hat{\lambda} - \lambda| : \lambda \in [\lambda_s, \lambda_r], \hat{\lambda} \in (-\infty, \hat{\lambda}_{s+1}] \cup [\hat{\lambda}_{r-1}, \infty)\}$, where we define $\hat{\lambda}_0 = -\infty$ and $\hat{\lambda}_{p+1} = \infty$, and assume that $\delta > 0$. Then*

$$\|\sin\Theta(\hat{V}, V)\|_{\mathrm{F}} \leq \frac{\|\hat{\Sigma} - \Sigma\|_{\mathrm{F}}}{\delta}. \tag{1}$$

Theorem 1 is an immediate consequence of Theorem V.3.6 of Stewart and Sun (1990). Despite the attractions of this bound, an obvious difficulty for statisticians is that we may have $\delta = 0$ for a particular realization of $\hat{\Sigma}$, even when the population eigenvalues are well-separated. As a toy example to illustrate this point, suppose that $\Sigma = \mathrm{diag}(50, 40, 30, 20, 10)$ and $\hat{\Sigma} = \mathrm{diag}(54, 37, 32, 23, 21)$. If we are interested in the eigenspaces spanned by the eigenvectors corresponding to the second, third and fourth largest eigenvalues, so $r = 2$ and $s = 4$, then Theorem 1 above cannot be applied, because $\delta = 0$.

Ignoring this issue for the moment, we remark that both occurrences of the Frobenius norm in (1) can be replaced with the operator norm $\|\cdot\|_{\mathrm{op}}$, or any other orthogonally invariant norm. Frequently in applications, we have $r = s = j$, say, in which case we can conclude that

$$\sin\Theta(\hat{v}_j, v_j) \leq \frac{\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}}{\min(|\hat{\lambda}_{j-1} - \lambda_j|, |\hat{\lambda}_{j+1} - \lambda_j|)}.$$

Since we may reverse the sign of $\hat{v}_j$ if necessary, there is a choice of orientation of $\hat{v}_j$ for which $\hat{v}_j^\top v_j \geq 0$. For this choice, we can also deduce that $\|\hat{v}_j - v_j\| \leq 2^{1/2}\sin\Theta(\hat{v}_j, v_j)$, where $\|\cdot\|$ denotes the Euclidean norm.

Theorem 1 is typically used to show that $\hat{v}_j$ is close to v_j as follows: first, we argue that $\hat{\Sigma}$ is close to Σ. This is often straightforward; for instance, when Σ is a population covariance matrix, it may be that $\hat{\Sigma}$ is just an empirical average of independent and identically distributed random matrices; cf. Section 3. Then we argue, e.g. using Weyl's inequality (Weyl, 1912; Stewart and Sun, 1990), that with high probability, $|\hat{\lambda}_{j-1} - \lambda_j| \geq (\lambda_{j-1} - \lambda_j)/2$ and $|\hat{\lambda}_{j+1} - \lambda_j| \geq (\lambda_j - \lambda_{j+1})/2$, so on these events $\|\hat{v}_j - v_j\|$ is small provided we are also willing to assume an eigenvalue separation, or eigen-gap, condition on the population eigenvalues.

The main contribution of this paper, in Theorem 2 in Section 2 below, is to give a variant of the Davis–Kahan $\sin\theta$ theorem that has two advantages for statisticians. First, the only eigen-gap condition is on the population eigenvalues, in contrast to the definition of δ in Theorem 1 above. Similarly, only population eigenvalues appear in the denominator of the bounds. This means there is no need for the statistician to worry about the event where $|\hat{\lambda}_{j-1} - \lambda_j|$ or $|\hat{\lambda}_{j+1} - \lambda_j|$ is small. Second, we show that the expression $\|\hat{\Sigma} - \Sigma\|_{\mathrm{F}}$ appearing in the numerator of the bound in (1) can be replaced with $\min(d^{1/2}\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}, \|\hat{\Sigma} - \Sigma\|_{\mathrm{F}})$. In Section 3, we give applications where our result could be used to allow authors to assume more natural conditions or to simplify

proofs, and also give a detailed example to illustrate the potential improvements of our bounds. The recent result of Vu et al. (2013, Corollary 3.1) has some overlap with our Theorem 2. We discuss the differences between our work and theirs shortly after the statement of Theorem 2.

Singular value decomposition, which may be regarded as a generalization of eigendecomposition, but which exists even when a matrix is not square, also plays an important role in many modern algorithms in statistics and machine learning. Examples include matrix completion (Candès and Recht, 2009), robust principal components analysis (Candès et al., 2009) and motion analysis (Kukush et al., 2002), among many others. Wedin (1972) provided the analogue of the Davis–Kahan $\sin\theta$ theorem for such general real matrices, working with singular vectors rather than eigenvectors, but with conditions and bounds that mix sample and population singular values. In Section 4, we extend the results of Section 2 to such settings; again our results depend only on a condition on the population singular values. Proofs are deferred to the Appendix.

2. MAIN RESULTS

THEOREM 2. *Let $\Sigma, \hat{\Sigma} \in \mathbb{R}^{p\times p}$ be symmetric, with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ and $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$ respectively. Fix $1 \leq r \leq s \leq p$ and assume that $\min(\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1}) > 0$, where we define $\lambda_0 = \infty$ and $\lambda_{p+1} = -\infty$. Let $d = s - r + 1$, and let $V = (v_r, v_{r+1}, \ldots, v_s) \in \mathbb{R}^{p\times d}$ and $\hat{V} = (\hat{v}_r, \hat{v}_{r+1}, \ldots, \hat{v}_s) \in \mathbb{R}^{p\times d}$ have orthonormal columns satisfying $\Sigma v_j = \lambda_j v_j$ and $\hat{\Sigma}\hat{v}_j = \hat{\lambda}_j \hat{v}_j$ for $j = r, r+1, \ldots, s$. Then*

$$\|\sin\Theta(\hat{V}, V)\|_{\mathrm{F}} \leq \frac{2\min(d^{1/2}\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}, \|\hat{\Sigma} - \Sigma\|_{\mathrm{F}})}{\min(\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1})}. \tag{2}$$

Moreover, there exists an orthogonal matrix $\hat{O} \in \mathbb{R}^{d\times d}$ such that

$$\|\hat{V}\hat{O} - V\|_{\mathrm{F}} \leq \frac{2^{3/2}\min(d^{1/2}\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}, \|\hat{\Sigma} - \Sigma\|_{\mathrm{F}})}{\min(\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1})}. \tag{3}$$

As mentioned briefly in the introduction, apart from the fact that we only impose a population eigen-gap condition, the main difference between this result and that given in Theorem 1 is in the $\min(d^{1/2}\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}, \|\hat{\Sigma} - \Sigma\|_{\mathrm{F}})$ term in the numerator of the bounds. In fact, the original statement of the Davis–Kahan $\sin\theta$ theorem has a numerator of $\|V\Lambda - \hat{\Sigma}V\|_{\mathrm{F}}$ in our notation, where $\Lambda = \mathrm{diag}(\lambda_r, \lambda_{r+1}, \ldots, \lambda_s)$. However, in order to apply that theorem in practice, statisticians have bounded this expression by $\|\hat{\Sigma} - \Sigma\|_{\mathrm{F}}$, yielding the bound in Theorem 1. When p is large, though, one would often anticipate that $\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}$, which is the ℓ_∞ norm of the vector of eigenvalues of $\hat{\Sigma} - \Sigma$, may well be much smaller than $\|\hat{\Sigma} - \Sigma\|_{\mathrm{F}}$, which is the ℓ_2 norm of this vector of eigenvalues. Thus when $d \ll p$, as will often be the case in practice, the minimum in the numerator may well be attained by the first term. It is immediately apparent from (A3) and (A4) in our proof that the smaller numerator $\|\hat{V}\Lambda - \Sigma\hat{V}\|_{\mathrm{F}}$ could also be used in our bound for $\|\sin\Theta(\hat{V}, V)\|_{\mathrm{F}}$ in Theorem 2, while $2^{1/2}\|\hat{V}\Lambda - \Sigma\hat{V}\|_{\mathrm{F}}$ could be used in our bound for $\|\hat{V}\hat{O} - V\|_{\mathrm{F}}$. Our reason for presenting the weaker bound in Theorem 2 is to aid direct applicability; see Section 3 for examples.

As mentioned in the introduction, Vu et al. (2013, Corollary 3.1) is similar in spirit to Theorem 2 above, and only involves a population eigen-gap condition, but there are some important differences. First, their result focuses on the eigenvectors corresponding to the top d eigenvalues, whereas ours applies to any set of d eigenvectors corresponding to a block of d consecutive eigenvalues, as in the original Davis–Kahan theorem. Their proof, which uses quite different

techniques from ours, does not appear to generalize immediately to this setting. Second, Corollary 3.1 of Vu et al. (2013) does not include the $d^{1/2}\|\hat{\Sigma}-\Sigma\|_{\mathrm{op}}$ term in the numerator of the bound. As discussed in the previous paragraph, it is this term that would typically be expected to attain the minimum in (2), especially in high-dimensional contexts. We also provide Theorem 3 to generalize the result to asymmetric or non-square matrices.

The constants presented in Theorem 2 are sharp, as the following example illustrates. Fix $d \in \{1, \ldots, \lfloor p/2 \rfloor\}$ and let $\Sigma = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$, where $\lambda_1 = \cdots = \lambda_{p-2d} = 5$, $\lambda_{p-2d+1} = \cdots = \lambda_{p-d} = 3$ and $\lambda_{p-d+1} = \cdots = \lambda_p = 1$. Suppose that $\hat{\Sigma}$ is also diagonal, with first $p-2d$ diagonal entries equal to 5, next d diagonal entries equal to 2, and last d diagonal entries equal to $2+\epsilon$, for some $\epsilon \in (0,3)$. If we are interested in the middle block of eigenvectors corresponding to those with corresponding eigenvalue 3 in Σ, then for every orthogonal matrix $\hat{O} \in \mathbb{R}^{d\times d}$,

$$\|\hat{V}\hat{O} - V\|_{\mathrm{F}} = 2^{1/2}\|\sin\Theta(\hat{V}, V)\|_{\mathrm{F}} = (2d)^{1/2} \leq (2d)^{1/2}(1+\epsilon) = \frac{2^{3/2}d^{1/2}\|\hat{\Sigma}-\Sigma\|_{\mathrm{op}}}{\min(\lambda_{p-2d}-\lambda_{p-2d+1}, \lambda_{p-d}-\lambda_{p-d+1})}.$$

In this example, the column spaces of V and $\hat{V}$ were orthogonal. However, even when these column spaces are close, our bound (2) is tight up to a factor of 2, while our bound (3) is tight up to a factor of $2^{3/2}$. To see this, suppose that $\Sigma = \mathrm{diag}(3,1)$ while $\hat{\Sigma} = \hat{V}\mathrm{diag}(3,1)\hat{V}^\top$, where

$$\hat{V} = \begin{pmatrix} (1-\epsilon^2)^{1/2} & -\epsilon \\ \epsilon & (1-\epsilon^2)^{1/2} \end{pmatrix}$$

for some $\epsilon > 0$. If $v = (1,0)^\top$ and $\hat{v} = \big((1-\epsilon^2)^{1/2}, -\epsilon\big)^\top$ denote the top eigenvectors of Σ and $\hat{\Sigma}$ respectively, then

$$\sin\Theta(\hat{v}, v) = \epsilon, \quad \|\hat{v}-v\|^2 = 2 - 2(1-\epsilon^2)^{1/2}, \quad \frac{2\|\hat{\Sigma}-\Sigma\|_{\mathrm{op}}}{3-1} = 2\epsilon.$$

Another theorem in Davis and Kahan (1970), the so-called $\sin 2\theta$ theorem, provides a bound for $\|\sin 2\Theta(\hat{V}, V)\|_{\mathrm{F}}$ assuming only a population eigen-gap condition. In the case $d = 1$, this quantity can be related to the square of the length of the difference between the sample and population eigenvectors $\hat{v}$ and v as follows:

$$\sin^2 2\Theta(\hat{v}, v) = (2\hat{v}^\top v)^2\{1 - (\hat{v}^\top v)^2\} = \frac{1}{4}\|\hat{v}-v\|^2(2-\|\hat{v}-v\|^2)(4-\|\hat{v}-v\|^2). \quad (4)$$

Equation (4) reveals, however, that $\|\sin 2\Theta(\hat{V}, V)\|_{\mathrm{F}}$ is unlikely to be of immediate interest to statisticians, and we are not aware of applications of the Davis–Kahan $\sin 2\theta$ theorem in statistics. No general bound for $\|\sin\Theta(\hat{V}, V)\|_{\mathrm{F}}$ or $\|\hat{V}\hat{O} - V\|_{\mathrm{F}}$ can be derived from the Davis–Kahan $\sin 2\theta$ theorem since we would require further information such as $\hat{v}^\top v \geq 1/2^{1/2}$ when $d = 1$, which would typically be unavailable. The utility of our bound comes from the fact that it provides direct control of the main quantities of interest to statisticians.

Many if not most applications of this result will only need $s = r$, i.e. $d = 1$. In that case, the statement simplifies a little; for ease of reference, we state it as a corollary:

COROLLARY 1. *Let $\Sigma, \hat{\Sigma} \in \mathbb{R}^{p\times p}$ be symmetric, with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ and $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$ respectively. Fix $j \in \{1, \ldots, p\}$, and assume that $\min(\lambda_{j-1} - \lambda_j, \lambda_j - \lambda_{j+1}) > 0$,*

where we define $\lambda_0 = \infty$ *and* $\lambda_{p+1} = -\infty$. *If* $v, \hat{v} \in \mathbb{R}^p$ *satisfy* $\Sigma v = \lambda_j v$ *and* $\hat{\Sigma}\hat{v} = \hat{\lambda}_j \hat{v}$, *then*

$$\sin\Theta(\hat{v}, v) \leq \frac{2\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}}{\min(\lambda_{j-1} - \lambda_j, \lambda_j - \lambda_{j+1})}.$$

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

$$\|\hat{v} - v\| \leq \frac{2^{3/2}\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}}{\min(\lambda_{j-1} - \lambda_j, \lambda_j - \lambda_{j+1})}.$$

3. Applications in statistical contexts

In the introduction, we explained how the fact that our variant of the Davis–Kahan $\sin\theta$ theorem only relies on a population eigen-gap condition can be used to simplify many arguments in the statistical literature. These include the work of Fan et al. (2013) on large covariance matrix estimation problems, Cai et al. (2013) on sparse principal component estimation, and an unpublished 2013 technical report by J. Fan and X. Han on estimating the false discovery proportion in large-scale multiple testing with highly correlated test statistics. Although our notation suggests that we have covariance matrix estimation in mind, we emphasize that the real, symmetric matrices in Theorem 2 are arbitrary, and could be for example inverse covariance matrices, or graph Laplacians as in the work of von Luxburg (2007) and Rohe et al. (2011) on spectral clustering in community detection with network data.

We now give some simple examples to illustrate the improvements afforded by our bound in Theorem 2. Consider the spiked covariance model in which $X_1, \ldots, X_n$ are independent random vectors having the $N_p(0, \Sigma)$ distribution, where $\Sigma = (\Sigma_{jk})$ is a diagonal matrix with $\Sigma_{jj} = 1 + \theta$ for some $\theta > 0$ for $1 \leq j \leq d$ and $\Sigma_{jj} = 1$ for $d + 1 \leq j \leq p$. Let $\hat{\Sigma} = n^{-1}\sum_{i=1}^n X_i X_i^\top$ denote the sample covariance matrix, and let V and $\hat{V}$ denote the matrices whose columns are unit-length eigenvectors corresponding to the d largest eigenvalues of Σ and $\hat{\Sigma}$ respectively. Fixing $n = 1000$, $p = 200$, $d = 10$ and $\theta = 1$, we found that our bound (2) from Theorem 2 was an improvement over that from (1) in every one of 100 independent data sets drawn from this model. In fact no bound could be obtained from Theorem 1 for 25 realizations because δ defined in that result was zero. The median value of $\|\sin\Theta(\hat{V}, V)\|_{\mathrm{F}}$ was 1·80, while the median values of the right-hand sides of (2) and (1) were 7·30 and 376 respectively. Some insight into the reasons for this marked improvement can be gained by considering an asymptotic regime in which $p/n \to \gamma \in (0, 1)$ as $n \to \infty$ and d and θ are considered fixed. Then, in the notation of Theorem 1,

$$\delta = \max(\lambda_d - \hat{\lambda}_{d+1}, 0) \to \max(\theta - 2\gamma^{1/2} - \gamma, 0),$$

almost surely, where the limit follows from Baik and Silverstein (2006, Theorem 1.1). On the other hand, the denominator of the right-hand side of (2) in Theorem 2 is θ, which may be much larger than $\max(\theta - 2\gamma^{1/2} - \gamma, 0)$. For the numerator, in this example, it can be shown that

$$E(\|\hat{\Sigma} - \Sigma\|_{\mathrm{F}}^2) = \frac{p(p+2)}{n} + \frac{2d(p+2)}{n}\theta + \frac{d(d+2)}{n}\theta^2 \geq \frac{p^2}{n}.$$

Moreover, by Theorem 1.1(b) of Baik, Ben Arous and Péché (2005) and a uniform integrability argument,

$$E(\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}^2) \leq E\{(\hat{\lambda}_1 - 1)^2\} \to \left\{\theta + \frac{(1+\theta)\gamma}{\theta}\right\}^2.$$

We therefore expect the minimum in the numerator of (2) to be attained by the term $d^{1/2}\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}$ in this example.

To illustrate our bound in a high-dimensional context, consider the data generating mechanism in our previous example. Given an even integer $k \in \{1, \ldots, p\}$, let $\hat{\Sigma} = \hat{\Sigma}_k$ be the tapering estimator for high-dimensional sparse covariance matrices introduced by Cai et al. (2010). In other words, $\hat{\Sigma}$ is the Hadamard product of the sample covariance matrix and a weight matrix $W = (w_{ij}) \in \mathbb{R}^{p\times p}$, where

$$w_{ij} = \begin{cases} 1, & |i-j| \le k/2, \\ 2 - \frac{2|i-j|}{k}, & k/2 < |i-j| < k, \\ 0, & \text{otherwise.} \end{cases}$$

To compare the bounds provided by Theorems 1 and 2, we drew 100 data sets from this model for each of the settings $n \in \{1000, 2000\}$, $p \in \{2000, 4000\}$, $d = 10$, $\theta = 1$ and $k = 20$. The bound (2) improved on that in (1) for every realisation in each setting; the medians of these bounds are presented in Table 1.

Table 1. *Median values of the bounds obtained from* (1) *and* (2)

n	p	RHS1	RHS2	n	p	RHS1	RHS2
1000	2000	12·1	2·65	1000	4000	17·3	2·69
2000	2000	7·20	1·92	2000	4000	10·2	1·90

RHS1, the bound obtained from (1); RHS2, the bound obtained from (2).

4. EXTENSION TO GENERAL REAL MATRICES

We now describe how the results of Section 2 can be extended to situations where the matrices under study may not be symmetric and may not even be square, and where interest is in controlling the principal angles between corresponding singular vectors.

THEOREM 3. *Suppose that $A, \hat{A} \in \mathbb{R}^{p\times q}$ have singular values $\sigma_1 \ge \cdots \ge \sigma_{\min(p,q)}$ and $\hat{\sigma}_1 \ge \cdots \ge \hat{\sigma}_{\min(p,q)}$ respectively. Fix $1 \le r \le s \le \mathrm{rank}(A)$ and assume that $\min(\sigma_{r-1}^2 - \sigma_r^2, \sigma_s^2 - \sigma_{s+1}^2) > 0$, where we define $\sigma_0^2 = \infty$ and $\sigma_{\mathrm{rank}(A)+1}^2 = -\infty$. Let $d = s - r + 1$, and let $V = (v_r, v_{r+1}, \ldots, v_s) \in \mathbb{R}^{q\times d}$ and $\hat{V} = (\hat{v}_r, \hat{v}_{r+1}, \ldots, \hat{v}_s) \in \mathbb{R}^{q\times d}$ have orthonormal columns satisfying $Av_j = \sigma_j u_j$ and $\hat{A}\hat{v}_j = \hat{\sigma}_j \hat{u}_j$ for $j = r, r+1, \ldots, s$. Then*

$$\|\sin\Theta(\hat{V}, V)\|_{\mathrm{F}} \le \frac{2(2\sigma_1 + \|\hat{A} - A\|_{\mathrm{op}}) \min(d^{1/2}\|\hat{A} - A\|_{\mathrm{op}}, \|\hat{A} - A\|_{\mathrm{F}})}{\min(\sigma_{r-1}^2 - \sigma_r^2, \sigma_s^2 - \sigma_{s+1}^2)}.$$

Moreover, there exists an orthogonal matrix $\hat{O} \in \mathbb{R}^{d\times d}$ such that

$$\|\hat{V}\hat{O} - V\|_{\mathrm{F}} \le \frac{2^{3/2}(2\sigma_1 + \|\hat{A} - A\|_{\mathrm{op}}) \min(d^{1/2}\|\hat{A} - A\|_{\mathrm{op}}, \|\hat{A} - A\|_{\mathrm{F}})}{\min(\sigma_{r-1}^2 - \sigma_r^2, \sigma_s^2 - \sigma_{s+1}^2)}.$$

Theorem 3 gives bounds on the proximity of the right singular vectors of Σ and $\hat{\Sigma}$. Identical bounds also hold if V and $\hat{V}$ are replaced with the matrices of left singular vectors U and $\hat{U}$, where $U = (u_r, u_{r+1}, \ldots, u_s) \in \mathbb{R}^{p\times d}$ and $\hat{U} = (\hat{u}_r, \hat{u}_{r+1}, \ldots, \hat{u}_s) \in \mathbb{R}^{p\times d}$ have orthonormal columns satisfying $A^\top u_j = \sigma_j v_j$ and $\hat{A}^\top \hat{u}_j = \hat{\sigma}_j \hat{v}_j$ for $j = r, r+1, \ldots, s$.

As mentioned in the introduction, Theorem 3 can be viewed as a variant of the generalized $\sin\theta$ theorem of Wedin (1972). Similar to the situation for symmetric matrices, there are many places in the statistical literature where Wedin's result has been used, but where we argue that Theorem 3 above would be a more natural result to which to appeal. Examples include the papers of Van Huffel and Vandewalle (1989) on the accuracy of least squares techniques, Anandkumar et al. (2014) on tensor decompositions for learning latent variable models, Shabalin and Nobel (2013) on recovering a low rank matrix from a noisy version and Sun and Zhang (2012) on matrix completion.

ACKNOWLEDGEMENTS

The first and third authors are supported by the third author's Engineering and Physical Sciences Research Council Early Career Fellowship. The second author is supported by a Benefactors' scholarship from St John's College, Cambridge. We are grateful for the anonymous reviewers' constructive comments, which helped to improve the paper.

APPENDIX

We first state an elementary lemma that will be useful in several places.

LEMMA A1. *Let $A \in \mathbb{R}^{m\times n}$, and let $U \in \mathbb{R}^{m\times p}$ and $W \in \mathbb{R}^{n\times q}$ both have orthonormal rows. Then*

$$\|U^\top A W\|_{\mathrm{F}} = \|A\|_{\mathrm{F}}.$$

If instead $U \in \mathbb{R}^{m\times p}$ and $W \in \mathbb{R}^{n\times q}$ both have orthonormal columns, then

$$\|U^\top A W\|_{\mathrm{F}} \leq \|A\|_{\mathrm{F}}.$$

Proof. For the first claim,

$$\|U^\top A W\|_{\mathrm{F}}^2 = \mathrm{tr}(U^\top A W W^\top A^\top U) = \mathrm{tr}(A A^\top U U^\top) = \mathrm{tr}(A A^\top) = \|A\|_{\mathrm{F}}^2.$$

For the second part, find a matrix $U_1 \in \mathbb{R}^{m\times(m-p)}$ such that $\begin{pmatrix} U & U_1 \end{pmatrix}$ is orthogonal, and a matrix $W_1 \in \mathbb{R}^{n\times(n-q)}$ such that $\begin{pmatrix} W & W_1 \end{pmatrix}$ is orthogonal. Then

$$\|A\|_{\mathrm{F}} = \left\| \begin{pmatrix} U^\top \\ U_1^\top \end{pmatrix} A \begin{pmatrix} W & W_1 \end{pmatrix} \right\|_{\mathrm{F}} \geq \left\| \begin{pmatrix} U^\top \\ U_1^\top \end{pmatrix} A W \right\|_{\mathrm{F}} \geq \|U^\top A W\|_{\mathrm{F}}.$$

Proof of Theorem 2. Let $\Lambda = \mathrm{diag}(\lambda_r, \lambda_{r+1}, \ldots, \lambda_s)$ and $\hat{\Lambda} = \mathrm{diag}(\hat{\lambda}_r, \hat{\lambda}_{r+1}, \ldots, \hat{\lambda}_s)$. Then

$$0 = \hat{\Sigma}\hat{V} - \hat{V}\hat{\Lambda} = \Sigma\hat{V} - \hat{V}\Lambda + (\hat{\Sigma} - \Sigma)\hat{V} - \hat{V}(\hat{\Lambda} - \Lambda).$$

Hence

$$\begin{aligned} \|\hat{V}\Lambda - \Sigma\hat{V}\|_{\mathrm{F}} &\leq \|(\hat{\Sigma} - \Sigma)\hat{V}\|_{\mathrm{F}} + \|\hat{V}(\hat{\Lambda} - \Lambda)\|_{\mathrm{F}} \\ &\leq d^{1/2}\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}} + \|\hat{\Lambda} - \Lambda\|_{\mathrm{F}} \leq 2d^{1/2}\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}, \end{aligned} \tag{A1}$$

where we have used Lemma 1 in the second inequality and Weyl's inequality (e.g. Stewart and Sun, 1990, Corollary IV.4.9) for the final bound. Alternatively, we can argue that

$$\begin{aligned} \|\hat{V}\Lambda - \Sigma\hat{V}\|_{\mathrm{F}} &\leq \|(\hat{\Sigma} - \Sigma)\hat{V}\|_{\mathrm{F}} + \|\hat{V}(\hat{\Lambda} - \Lambda)\|_{\mathrm{F}} \\ &\leq \|\hat{\Sigma} - \Sigma\|_{\mathrm{F}} + \|\hat{\Lambda} - \Lambda\|_{\mathrm{F}} \leq 2\|\hat{\Sigma} - \Sigma\|_{\mathrm{F}}, \end{aligned} \tag{A2}$$

where the second inequality follows from two applications of Lemma 1, and the final inequality follows from the Wielandt–Hoffman theorem (e.g. Wilkinson, 1965, pp. 104–8).

Let $\Lambda_1 = \mathrm{diag}(\lambda_1, \ldots, \lambda_{r-1}, \lambda_{s+1}, \ldots, \lambda_p)$, and let V_1 be a $p \times (p-d)$ matrix such that $P = \begin{pmatrix} V & V_1 \end{pmatrix}$ is orthogonal and such that

$$P^\top \Sigma P = \begin{pmatrix} \Lambda & 0 \\ 0 & \Lambda_1 \end{pmatrix}.$$

Then

$$\begin{aligned} \|\hat{V}\Lambda - \Sigma\hat{V}\|_{\mathrm{F}} &= \|VV^\top\hat{V}\Lambda + V_1V_1^\top\hat{V}\Lambda - V\Lambda V^\top\hat{V} - V_1\Lambda_1V_1^\top\hat{V}\|_{\mathrm{F}} \\ &\geq \|V_1V_1^\top\hat{V}\Lambda - V_1\Lambda_1V_1^\top\hat{V}\|_{\mathrm{F}} \geq \|V_1^\top\hat{V}\Lambda - \Lambda_1V_1^\top\hat{V}\|_{\mathrm{F}}, \end{aligned} \tag{A3}$$

where the first inequality follows because $V^\top V_1 = 0$, and the second from another application of Lemma 1. For real matrices A and B, we write $A \otimes B$ for their Kronecker product (e.g. Stewart and Sun, 1990, p. 30) and $\mathrm{vec}(A)$ for the vectorization of A, i.e. the vector formed by stacking its columns. We recall the standard identity $\mathrm{vec}(ABC) = (C^\top \otimes A)\mathrm{vec}(B)$, which holds whenever the dimensions of the matrices are such that the matrix multiplication is well-defined. We also write I_m for the m-dimensional identity matrix. Then

$$\begin{aligned} \|V_1^\top\hat{V}\Lambda - \Lambda_1V_1^\top\hat{V}\|_{\mathrm{F}} &= \|(\Lambda \otimes I_{p-d} - I_d \otimes \Lambda_1)\mathrm{vec}(V_1^\top\hat{V})\| \\ &\geq \min(\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1})\|\mathrm{vec}(V_1^\top\hat{V})\| \\ &= \min(\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1})\|\sin\Theta(\hat{V}, V)\|_{\mathrm{F}}, \end{aligned} \tag{A4}$$

since

$$\begin{aligned} \|\mathrm{vec}(V_1^\top\hat{V})\|^2 = \mathrm{tr}(\hat{V}^\top V_1V_1^\top\hat{V}) = \mathrm{tr}\big\{(I_p - VV^\top)\hat{V}\hat{V}^\top\big\} &= d - \|\hat{V}^\top V\|_{\mathrm{F}}^2 \\ &= \|\sin\Theta(\hat{V}, V)\|_{\mathrm{F}}^2. \end{aligned}$$

We deduce from (A4), (A3), (A2) and (A1) that

$$\|\sin\Theta(\hat{V}, V)\|_{\mathrm{F}} \leq \frac{\|V_1^\top\hat{V}\Lambda - \Lambda_1V_1^\top\hat{V}\|_{\mathrm{F}}}{\min(\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1})} \leq \frac{2\min(d^{1/2}\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}}, \|\hat{\Sigma} - \Sigma\|_{\mathrm{F}})}{\min(\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1})},$$

as required.

For the second conclusion, by a singular value decomposition, we can find orthogonal matrices $\hat{O}_1, \hat{O}_2 \in \mathbb{R}^{d\times d}$ such that $\hat{O}_1^\top\hat{V}^\top V\hat{O}_2 = \mathrm{diag}(\cos\theta_1, \ldots, \cos\theta_d)$, where $\theta_1, \ldots, \theta_d$ are the principal angles between the column spaces of V and $\hat{V}$. Setting $\hat{O} = \hat{O}_1\hat{O}_2^\top$, we have

$$\begin{aligned} \|\hat{V}\hat{O} - V\|_{\mathrm{F}}^2 = \mathrm{tr}\big\{(\hat{V}\hat{O} - V)^\top(\hat{V}\hat{O} - V)\big\} &= 2d - 2\mathrm{tr}(\hat{O}_2\hat{O}_1^\top\hat{V}^\top V) \\ &= 2d - 2\sum_{j=1}^d \cos\theta_j \leq 2d - 2\sum_{j=1}^d \cos^2\theta_j = 2\|\sin\Theta(\hat{V}, V)\|_{\mathrm{F}}^2. \end{aligned} \tag{A5}$$

The result now follows from our first conclusion. □

Proof of Theorem 3. The matrices $A^\top A, \hat{A}^\top\hat{A} \in \mathbb{R}^{q\times q}$ are symmetric, with eigenvalues $\sigma_1^2 \geq \cdots \geq \sigma_q^2$ and $\hat{\sigma}_1^2 \geq \cdots \geq \hat{\sigma}_q^2$ respectively. Moreover, we have $A^\top Av_j = \sigma_j^2 v_j$ and $\hat{A}^\top\hat{A}\hat{v}_j = \hat{\sigma}_j^2\hat{v}_j$ for $j = r, r+1, \ldots, s$. We deduce from Theorem 2 that

$$\|\sin\Theta(\hat{V}, V)\|_{\mathrm{F}} \leq \frac{2\min(d^{1/2}\|\hat{A}^\top\hat{A} - A^\top A\|_{\mathrm{op}}, \|\hat{A}^\top\hat{A} - A^\top A\|_{\mathrm{F}})}{\min(\sigma_{r-1}^2 - \sigma_r^2, \sigma_s^2 - \sigma_{s+1}^2)}. \tag{A6}$$

Now, by the submultiplicity of the operator norm,

$$\begin{aligned} \|\hat{A}^\top\hat{A} - A^\top A\|_{\mathrm{op}} = \|(\hat{A} - A)^\top\hat{A} + A^\top(\hat{A} - A)\|_{\mathrm{op}} &\leq (\|\hat{A}\|_{\mathrm{op}} + \|A\|_{\mathrm{op}})\|\hat{A} - A\|_{\mathrm{op}} \\ &\leq (2\sigma_1 + \|\hat{A} - A\|_{\mathrm{op}})\|\hat{A} - A\|_{\mathrm{op}}. \end{aligned} \tag{A7}$$

On the other hand,

$$\begin{aligned}\|\hat{A}^\top\hat{A} - A^\top A\|_{\mathrm{F}} &= \|(\hat{A}-A)^\top\hat{A} + A^\top(\hat{A}-A)\|_{\mathrm{F}} \\ &\le \|(\hat{A}^\top \otimes I_q)\mathrm{vec}\big((\hat{A}-A)^\top\big)\| + \|(I_p \otimes A^\top)\mathrm{vec}(\hat{A}-A)\| \\ &\le (\|\hat{A}^\top \otimes I_q\|_{\mathrm{op}} + \|I_p \otimes A^\top\|_{\mathrm{op}})\|\hat{A}-A\|_{\mathrm{F}} \\ &\le (2\sigma_1 + \|\hat{A}-A\|_{\mathrm{op}})\|\hat{A}-A\|_{\mathrm{F}}. \qquad \text{(A8)}\end{aligned}$$

We deduce from (A6), (A7) and (A8) that

$$\|\sin\Theta(\hat{V},V)\|_{\mathrm{F}} \le \frac{2(2\sigma_1 + \|\hat{A}-A\|_{\mathrm{op}})\min(d^{1/2}\|\hat{A}-A\|_{\mathrm{op}}, \|\hat{A}-A\|_{\mathrm{F}})}{\min(\sigma_{r-1}^2 - \sigma_r^2, \sigma_s^2 - \sigma_{s+1}^2)}.$$

The bound for $\|\hat{V}\hat{O} - V\|_{\mathrm{F}}$ now follows immediately from this and (A5). □

REFERENCES

ANANDKUMAR, A., GE. R., HSU, D., KAKADE, S. M. & TELGARSKY, M. (2014). Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.* **15**, 2773–832.

BAIK, J., BEN AROUS, G. & PÉCHÉ, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.* **33**, 1643–97.

BAIK, J. & SILVERSTEIN, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *J. Multivariate Anal.* **97**, 1382–408.

CAI, T. T., MA, Z. & WU, Y. (2013). Sparse PCA: Optimal rates and adaptive estimation. *Ann. Statist.* **41**, 3074–110.

CAI, T. T., ZHANG, C.-H. & ZHOU, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38**, 2118–44.

CANDÈS, E. J., LI, X., MA, Y. & WRIGHT, J. (2009). Robust Principal Component Analysis? *Journal of the ACM* **58**, 1–37.

CANDÈS E. J. & RECHT, B. (2009). Exact matrix completion via convex optimization. *Found. of Comput. Math.* **9**, 717–72.

DAVIS, C. & KAHAN, W. M. (1970). The rotation of eigenvectors by a pertubation. III. *SIAM J. Numer. Anal.* **7**, 1–46.

DONATH, W. E. & HOFFMAN, A. J. (1973). Lower bounds for the partitioning of graphs. *IBM Journal of Research and Development* **17**, 420–5.

FAN, J., LIAO, Y. & MINCHEVA, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *J. Roy. Statist. Soc., Ser. B* **75**, 603–80.

KUKUSH, A., MARKOVSKY, I. & VAN HUFFEL, S. (2002). Consistent fundamental matrix estimation in a quadratic measurement error model arising in motion analysis. *Comput. Stat. Data An.* **41**, 3–18.

ROHE, K., CHATTERJEE, S. & YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39**, 1878–915.

SHABALIN, A. A. & NOBEL, A. B. (2013). Reconstruction of a low-rank matrix in the presence of Gaussian noise. *J. Mult. Anal* **118**, 67–76.

STEWART, G. W. & SUN, J.-G. (1990). *Matrix Perturbation Theory*. Academic Press, Inc.: San Diego, California.

SUN T. & ZHANG, C.-H. (2012). Calibrated elastic regularization in matrix completion. *Adv. Neural Inf. Proc. Sys.* **25**, Ed. F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger. MIT Press: Cambridge, Massachusetts.

VAN HUFFEL, S. & VANDEWALLE, J. (1989). On the accuracy of total least squares and least squares techniques in the presence of errors on all data. *Automatica* **25**, 765–9.

VON LUXBURG, U. (2007). A tutorial on spectral clustering. *Statist. Comput.* **17**, 395–416.

VU, V. Q., CHO, J., LEI, J. & ROHE, K. (2013). Fantope projection and selection: a near-optimal convex relaxation of sparse PCA. *Adv. Neural Inf. Proc. Sys.* **26**, Ed. C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger. Nips Foundation.

WEDIN, P.-Å. (1972). Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics* **12**, 99–111.

WEYL, H. (1912). Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf der Theorie der Hohlraumstrahlung). *Mathematische Annalen* **71**, 441–79.

WILKINSON, J. H. (1965). *The Algebraic Eigenvalue Problem*. Clarendon Press, Oxford.

ZOU, H., HASTIE, T. J. & TIBSHIRANI, R. J. (2006). Sparse Principal Components Analysis. *J. Comput. Graph. Statist.* **15**, 265–86.

[*Received May* 2014. *Revised May* 2014]